Statistical Modeling Method for Efficiency Improvement of Industrial Processes(Dissertation_全文)

AUTHOR(S):

Kim, Sanghong

Statistical Modeling Method for Efficiency Improvement of Industrial Processes

Sanghong Kim

2014

Contents

Nomenclature

Latin alphabets

d_n	: Distance beween a query $\boldsymbol{x}_q$ and the nth sample $\boldsymbol{x}_n$
M	: Number of input variables
N	: Number of samples
R	: Number of latent variables of a partial least square model
x_{nm}	: nth component of mth column of an input variable matrix $\boldsymbol{X}$
$\boldsymbol{x}_n \in \mathfrak{R}^M$	: nth row of an input variable matrix $\boldsymbol{X}$
$\boldsymbol{x}_q \in \mathfrak{R}^M$	: Query, for which an output estimation is required
$\boldsymbol{X} \in \mathfrak{R}^{N \times M}$	: Input variable matrix
$\tilde{\boldsymbol{X}} \in \mathfrak{R}^{N \times M}$	: Input variable matrix after data preprocessing
$\boldsymbol{y} \in \mathfrak{R}^N$	: Output variable vector

Greek alphabets

$\boldsymbol{\beta} \in \mathfrak{R}^M$	: Regression coefficient vector
$\boldsymbol{\Theta} = \mathrm{diag}(\theta_1, \theta_2, \cdots, \theta_M) \in \mathfrak{R}^{M \times M}$	: Weighting matrix
$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_M) \in \mathfrak{R}^{M \times M}$	: Input scaling factor matrix
φ	: Localization parameter
$\boldsymbol{\Omega} = \mathrm{diag}(\omega_1, \omega_2, \cdots, \omega_N) \in \mathfrak{R}^{N \times N}$	: Similarity matrix

Chapter 1

Introduction

1.1 Background

In process industries, one of the most important tasks is to ensure product quality and to reduce operation cost to keep competitiveness in a global market. However, real-time measurement of product quality is not always available because of high measurement equipment cost and long measurement time. To solve this problem, many researches on soft-sensors, which estimate product quality by using real-time measurements, have been conducted [1–3]. By using a soft-sensor, inferential control based on output estimates can be realized, and operation cost can be reduced. To develop a soft-sensor, a white-box model (1st principle model), which is based on physical and chemical knowledge of processes, a black-box model (statistical model), which is based on statistical analysis of the process data, and a gray-box model, which is the combination of a white-box model and a black-box model, can be used. This paper focuses on soft-sensors using black-box models since white-box models are often too difficult to construct for complex industrial processes. Soft-sensors have already been used in many kinds of industries such as chemical/petrochemical, steel, semiconductor, pharmaceutical, and food industries. According to a questionnaire survey [3], in 2009 over 400 soft-sensors were working in

distillation and chemical reaction processes at 15 companies in Japan which answered the survey. However, some problems still remain to be solved as explained in the next section, and this research tries to solve some of these problems.

1.2 Soft-sensor design method

In this section, a general procedure of soft-sensor design is explained and the past researches on soft-sensor design are summarized.

1.2.1 General procedure

A general procedure of soft-sensor design is described in Figure 1.1. The first step of the off-line procedure is preparation of process data. In general, daily operation data is available while design of experiments (DoE) can be used in some cases. Next, data preprocessing and abnormal data detection are conducted to make the following analysis easier. Then, input variables which are correlated to an output variable are selected, and a soft-senor is constructed by using selected input variables and an output variable. The soft-sensor is validated and implemented in a real process if it is confirmed that the estimation accuracy of the soft-sensor is high enough. If the estimation accuracy is not satisfactory, the above steps are repeated. In the on-line procedure, a query, for which an output estimation is required, is obtained and the output estimate is calculated by using the soft-sensor. When an output measurement becomes available, the soft-sensor can be updated if needed. The past researches on each step are summarized in the following sections.

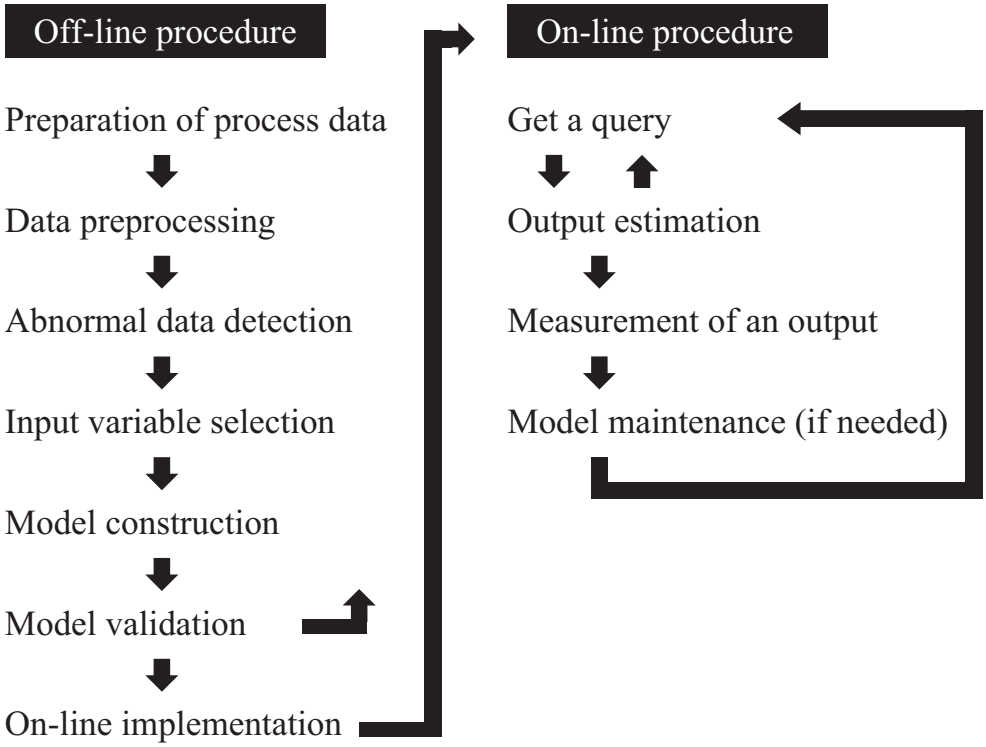

Figure 1.1: General procedure of soft-sensor design

1.2.2 Data preprocessing

Data preprocessing methods can be divide into 4 groups: centering, smoothing, input scaling and the others such as Fourier transformation. Figure 1.2 shows the concept of centering, smoothing and input scaling.

Centering is conducted by subtracting the mean value of each input variable as shown in the following equations.

$$\tilde{\boldsymbol{X}} = \boldsymbol{X} - \mathbf{1}_N[\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_M] \tag{1.1}$$

$$\bar{x}_m = \frac{1}{N}\sum_{n=1}^{N} x_{nm} \qquad (m = 1, 2, \cdots, M) \tag{1.2}$$

Here, $\boldsymbol{X} \in \Re^{N\times M}$ is an input variable matrix before centering, $\tilde{\boldsymbol{X}} \in \Re^{N\times M}$ is an input variable matrix after centering, $\mathbf{1}_N \in \Re^N$ is a vector of ones, M is the number of input

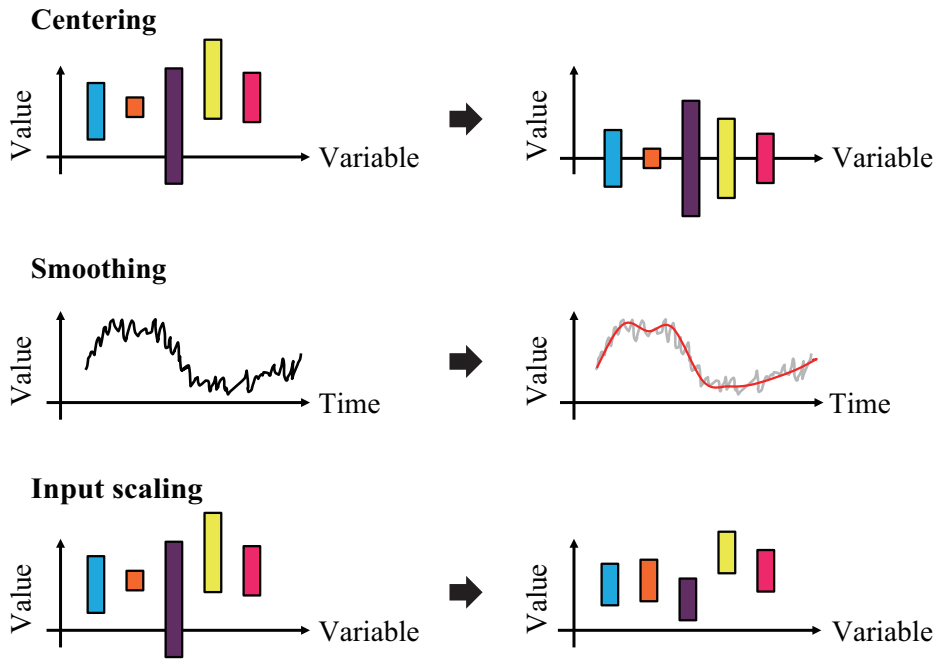

Figure 1.2: Concept of data preprocessing

variables, N is the number of samples, and x_{nm} is the nth component of the mth column of $\boldsymbol{X}$.

Centering is conducted to simplify equations for soft-sensor design. Hereafter, it is assumed that mean of every input variable in $\boldsymbol{X}$ is zero without loss of generality.

Smoothing is for reducing measurement noise of time series data. Moving average is a simple method for smoothing represented as follows:

$$\tilde{x}_{nm} = \sum_{\acute{n}=n-n_{av}}^{n} x_{\acute{n}m} \tag{1.3}$$

where $\tilde{x}_{nm}$ is smoothed data and n_{av} is the number of samples for smoothing.

Scaling is represented as

$$\tilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{\Lambda} \tag{1.4}$$

$$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_M) \tag{1.5}$$

where λ_m is a non-negative input scaling factor for the mth input variable. The most popular input scaling method is autoscale, in which λ_m is defined by the reciprocal of the standard deviation of the mth input variable [4, 5]. The effect of the difference of measurement units can be compensated by using autoscale, however, it is not always optimal in terms of the estimation accuracy of soft-sensors. The input scaling affects important statistical properties of the data such as distance between samples and covariance of samples, and it also affects the estimation result. However, researches on input variable scaling have not been actively conducted.

The other data preprocessing methods are used according to the property of process data. For example, Fourier transformation might be used for analysis of spectrum data such as sound, and wavelet transformation might be suitable for time-series data analysis.

1.2.3 Abnormal data detection

Process data usually have missing values and outlier because of sensor maintenance and malfunction. Since abnormal values can deteriorate the performance of a soft-sensor, they should be detected and removed. Shewhart chart [6] and Hampel identifier [7] are famous and easy-to-use outlier detection tools and have been adopted in practice. However, they check individual variables independently and their performance is not very high when variables are correlated to each other. Multivariate statistical process control (MSPC), based on principal component analysis (PCA) and partial least squares (PLS), was developed to solve this problem [8]. In MSPC, a subspace of the original input variable data is determined by PCA and PLS, then samples which are far from the subspace or the origin of subspace is regarded as outliers. Kamohara et al. [9] applied the PLS-based MSPC method to the ethylene fractionator at the Showa Denko K.K. Oita plant. Their method aimed to detect an outlier by monitoring Hotelling's T^2 and Q statistics. Independent component analysis (ICA) has been also adopted for detecting abnormal data. Kano et

al. [10] compared an ICA-based method and PCA-based MSPC in a numerical example and simulated CSTR process, and concluded that the ICA-based method was superior to PCA-based MSPC. More detailed information on statistical process monitoring methods for fault detection, identification and reconstruction is summarized by Qin [11].

1.2.4 Input variable selection

In general, some input variables are correlated to an output variable, and others are irrelevant to it. In addition, including input variables which are irrelevant to an output variable has bad influence on output estimation. Thus, it is crucial to select appropriate input variables to enhance the estimation accuracy, but the variable selection is regarded as one of the most difficult parts concerning soft-sensor development [12]. In many cases, experienced engineers select the input variables mainly based on their own process knowledge. However, it is time consuming for the engineers to select the input variables since trial and error is inevitable. In addition, the selected variable might not be optimal in terms of estimation accuracy. Also, it becomes very difficult even for experienced engineers to properly select input variables when the numbers of measured variables are large, and physical and chemical phenomena are not sufficiently understood. Thus, a systematic method for input variable selection is required to improve the estimation performance of soft-sensors and shorten the development period.

In the past, many indexes for selecting a set of input variables were proposed. A well-known index for multiple linear regression (MLR) is the F-value based on the statistical hypothesis test. Other popular indexes include adjusted coefficient of determination R^2, Akaike information criterion (AIC) [13], Mallow's C_p [14] and root mean square error of cross validation (RMSECV). These indexes evaluate a given set of input variables based on the fitness and the complexity of the corresponding regression model. However, they cannot select a set of input variables, and it is not practical to calculate an index for all

combinations of input variables when a large number of variables are measured. To select input variables efficiently, two types of methods are available: one uses optimization techniques to evaluate the goodness of combinations, and the other evaluates the importance of each input variable separately.

Among the optimization-based methods are the stepwise method and genetic algorithm (GA). The stepwise method repeatedly constructs models by adding or removing a variable with a greedy algorithm. GA is an algorithm that mimics the process of natural evolution. GA has been used to solve input variable selection problems formulated as mixed-integer problems [15, 16].

The indexes for evaluating each variable include magnitude of regression coefficient, variable influence on projection (VIP) [17], uninformative variable elimination (UVE) [18] and others. In this type of approach, input variables are selected if the corresponding indexes are larger than a threshold.

The idea of using the magnitude of regression coefficients for input variable selection is simple; variables having larger coefficients are more important. In this direction, an interesting method is least absolute shrinkage and selection operator (Lasso) [19], which is MLR with a penalty on the L-1 norm of the regression coefficient vector $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta}_{\text{Lasso}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \alpha\|\boldsymbol{\beta}\|_1 \tag{1.6}$$

where $\boldsymbol{y} \in \mathfrak{R}^N$ is an output variable vector and α is a tuning parameter. By penalizing the L-1 norm, unlike ridge regression that penalizes the L-2 norm, Lasso forces some regression coefficients to be zero; consequently the corresponding input variables can be removed. The regression coefficient vector $\boldsymbol{\beta}_{\text{PLS}}$ derived by PLS is also applicable. The input variable selection method using $\boldsymbol{\beta}_{\text{PLS}}$ is referred to as PLS-Beta [20]. Another index for PLS is VIP [17]. The VIP score evaluates the influence of individual input variables on output estimation. Chong and Jun [20] applied the stepwise method, Lasso, PLS-Beta and VIP to artificial data and compared their performances. VIP and PLS-Beta

achieved the similar performance and were superior to the stepwise method and Lasso. UVE, proposed by Centner et al. [18], uses regression coefficients and their standard deviation for each input variable derived by leave-one-out cross validation (LOOCV). In other input variable selection methods, groups of input variables are generated and group-wise selection is conducted. For example, when spectrum data is analyzed, interval selection methods such as interval PLS (iPLS) can be used to make the groups of the input variables (wavelengths) [21]. In interval selection methods, neighboring wavelengths are grouped into one group since they are expected to have a similar effect on the output from the viewpoint of spectroscopy. Then, a selection index such as RMSECV or an optimization-based method is used to select the groups. Fujiwara et al. [22] proposed a novel grouping method based on the correlation between input variables. In their method, nearest correlation spectral clustering (NCSC) [23, 24] is used for variable grouping. They compared the proposed method with the conventional methods such as the stepwise method, PLS-Beta, VIP, Lasso and manual selection by the engineers with industrial operation data of an ethylene fractionator. The NCSC-based variable selection (NCSC-VS) outperformed the others. This result demonstrates the advantage of the group-wise variable selection method over the conventional methods. In addition, the computational load can be reduced by grouping input variables.

Other indexes can be found in literatures [3, 21, 25, 26]. Nevertheless, trial and error is unavoidable in practice because different methods can derive the best performance depending on the case [27], and it is not clearly understood how to choose an input variable selection method. Thus it is not enough to propose a new method and evaluate it in particular cases. The reason why a method functions well or not should also be investigated so that an appropriate input variable selection method can be found with less effort and hence the burden of soft-sensor development can be reduced.

1.2.5 Model construction

According to a questionnaire survey [3], MLR is the most frequently used method for model construction in practice, in which regression coefficient vector $\boldsymbol{\beta}_{\mathrm{MLR}}$ is defined as

$$\boldsymbol{\beta}_{\mathrm{MLR}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \,. \tag{1.7}$$

Though MLR has been widely accepted in practice, it has a problem that $\boldsymbol{\beta}_{\mathrm{MLR}}$ becomes unstable when input variables are correlated to each other: this problem is called multicollinearity problem. PLS can solve this problem by deriving uncorrelated input variables, which are linear combinations of original input variables. The algorithm of PLS is as follows:

1. Determine the number of latent variables R and set r to 1.

2. Calculate $\boldsymbol{X}_r$ and $\boldsymbol{y}_r$.

$$\boldsymbol{X}_r = \boldsymbol{X} \tag{1.8}$$

$$\boldsymbol{y}_r = \boldsymbol{y} \tag{1.9}$$

3. Derive the r-th latent variable of $\boldsymbol{X}$

$$\boldsymbol{t}_r = \boldsymbol{X}_r\boldsymbol{w}_r \tag{1.10}$$

   where $\boldsymbol{w}_r$ is the eigenvector of $\boldsymbol{X}_r^{\mathrm{T}}\boldsymbol{y}_r\boldsymbol{y}_r^{\mathrm{T}}\boldsymbol{X}_r$ which corresponds to the maximum eigen value.

4. Derive the r-th loading vector of $\boldsymbol{X}$

$$\boldsymbol{p}_r = \frac{\boldsymbol{X}_r^{\mathrm{T}}\boldsymbol{t}_r}{\boldsymbol{t}_r^{\mathrm{T}}\boldsymbol{t}_r} \tag{1.11}$$

   and the regression coefficient vector

$$\boldsymbol{q}_r = \frac{\boldsymbol{y}_r^{\mathrm{T}}\boldsymbol{t}_r}{\boldsymbol{t}_r^{\mathrm{T}}\boldsymbol{t}_r} \,. \tag{1.12}$$

5. If $r = R$, finish modeling. Otherwise, set

$$\boldsymbol{X}_{r+1} = \boldsymbol{X}_r - \boldsymbol{t}_r \boldsymbol{p}_r^{\mathrm{T}} \tag{1.13}$$

$$\boldsymbol{y}_{r+1} = \boldsymbol{y}_r - \boldsymbol{t}_r \boldsymbol{q}_r^{\mathrm{T}} \tag{1.14}$$

6. Set r to $r + 1$ and go to step 3.

PLS has been also used in practice as well as MLR [28–35], and this fact shows that linear models are practically useful. In some cases, however, nonlinear models are required to achieve high estimation accuracy for processes having strong nonlinearity. Thus, nonlinear modeling methods such as neural networks [36–39], support vector regression [40–42] and polynomial functions [43–45] have been used to construct nonlinear soft-sensors.

1.2.6 Model maintenance

When process characteristics and operation condition change after model validation, the estimation performance of a soft-sensor deteriorates and thus model maintenance, which takes long time and costs much, is required to keep the estimation performance. This problem is recognized as one of the most serious problems of soft-sensor design and application [3, 46, 47].

Recursive and just-in-time model

To reduce the burden of model maintenance, several recursive modeling methods, which update models by prioritizing newer samples, have been developed [48]. When process characteristics change gradually, the prioritized samples are supposed to be similar to a query. For such a case, the recursive methods can cope with gradual changes in process characteristics. However, they cannot cope with an abrupt change in process characteristics caused by replacement of a catalyst, cleaning of equipment, etc., because a query

sampled just after an abrupt change becomes significantly different from the prioritized samples.

Locally weighted regression (LWR) [49], which is also called just-in-time (JIT) learning, lazy learning or model-on-demand, constructs a local model by prioritizing samples in a database according to the similarity between them and a query. Hence, LWR can cope with abrupt changes as well as gradual ones in contrast to the recursive methods introduced in Kadlec et al. [48]. Furthermore, it can cope with nonlinearity. To build an accurate model with LWR, the similarity needs to be properly defined. In general, the similarity is defined on the basis of the Euclidean distance or the Mahalanobis distance [42,50–55]. Other similarity measures proposed so far include the angle [47,56,57], the distance between an output estimate for a query derived by a global model and output measurements for samples in a database [58], the correlation [23, 24] and the weighted Euclidean distance [59–61].

Database management

In addition to define the similarity properly, it is crucial to update a database when new data becomes available in order to cope with changes in process characteristics. However, researches on database management have not been actively conducted.

In general, the age and the density of samples are important indexes to evaluate the goodness of a database, since the estimation performance may deteriorate when the samples in the database are old and sparse. However, the relative importance of these indexes changes according to the nonlinearity and time-variance of processes, and also characteristic of changes in input variables. Figure 1.3 summarizes the importance of the indexes for different kinds of processes and the characteristic of changes in input variables. Here, processes are classified into four groups: linear time-invariant (LTIV), linear time-variant (LTV), nonlinear time-invariant (NLTIV) and nonlinear time-variant (NLTV). In addition, changes in input variables are classified into two types: gradual and abrupt. When

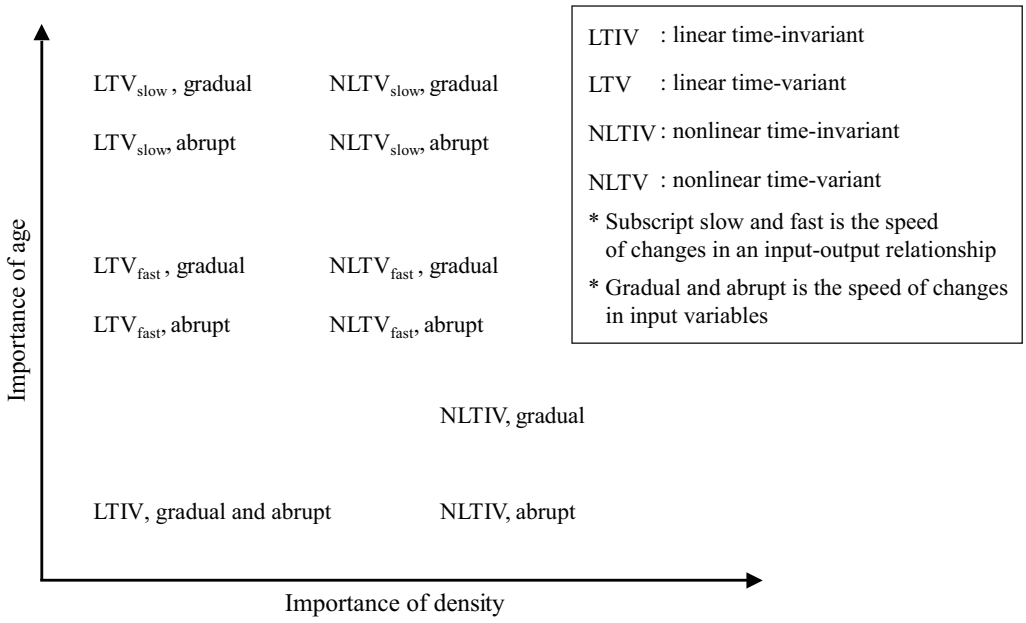

Figure 1.3: The importance of the age and the density of samples for database management

a process is LTIV, the high estimation performance will be achieved even if samples are old and sparse. When a process is LTV, the estimation performance will be improved by storing newer samples in the database since older samples cannot represent the current input-output relationship. In addition, the improvement becomes more significant as the changes in the input-output relationship becomes slower, or the changes in input variables becomes more gradual. When a process is NLTIV, the density of the samples is dominant; the higher density is required as the nonlinearity becomes stronger. When input variables change gradually, using the newer samples may improve the estimation performance since a query can be assumed to be similar to newer samples. When a process is NTLV, both the age and the density of samples are important.

A simple updating method is the moving time-window method, in which a newly obtained sample is added to the database and the oldest sample is removed from the database. This method has been used successfully in the steel process for more than

seven years [59]. In this application, regression coefficients of a global MLR model were used to determine the similarity used in LWR, and the estimation performance was improved. This result suggests that the process nonlinearity is not strong. In addition, the operating conditions are frequently changed to manufacture the various products for various customers. Also, the optimal condition for each product changes with time, therefore laborious table maintenance was required in the past. These results suggest that the process is $\mathrm{LTV}_{\mathrm{slow}}$ with abrupt input changes. Since the age of samples in the database is important for this process as shown in Figure 1.3, the moving time-window method is suitable for this application.

Wu et al. [62] proposed a database management method which takes into account both the age and the density of samples. When a new sample is added to a database, their method applies a clustering method based on the adaptive-resonance-theory-2 (ART2) or the WED between samples. Then, it checks the number of samples in the cluster whose sample size is the maximum. If the number is less than three, no sample is removed. Otherwise, the oldest sample in the cluster is removed. This method was compared to the moving time-window method through their applications to photo processes of a 5th-generation thin-film transistor liquid crystal display (TFT-LCD) factory. They recommended the use of the WED-based method in terms of estimation accuracy and the computational time. This result implies that the processes are NLTV and the density of samples is important to improve the estimation performance. Such characteristics are common in semiconductor processes since replacement and cleaning of equipment are often conducted.

As explained above, the age and the density of samples are important indexes for database management. To develop a better database management method, it seems crucial to evaluate the strength of the process nonlinearity to determine the density needed.

1.3 Thesis overview

This research aims to develop practically useful methods for improving the estimation accuracy of soft-sensors. In chapter 2, two input scaling methods are proposed to improve the accuracy of soft-sensors. The proposed methods can determine the scales of input variables based on their importance for soft-sensor design and reduce the estimation error. In chapter 3, an input variable selection method is proposed for batch processes. While it is applicable only to batch processes, it is quite easy to understand, and it can improve the estimation accuracy since it utilizes a specific property of batch processes. Chapter 4 shows successful application results of JIT model to real chemical processes. In addition, a new similarity is proposed for improving the estimation accuracy of JIT models. Finally, chapter 5 concludes this research and suggests the future direction of the researches on soft-sensor design.

Chapter 2

Input Variable Scaling Based on Importance

Abstract

The input variable scaling is one of the most important steps of soft-sensor design, however, it has not been actively investigated so far; autoscale is utilized in most cases. This research proposes two input variable scaling methods for improving the accuracy of soft-sensors. One of the proposed methods statistically derives the input scaling factors. The other utilizes spectroscopic data of a material whose content is an estimation target. The proposed methods can determine the scales of input variables based on their importance for soft-sensor design. Thus, bad influence of input variables which are not related to an output variable can be reduced, and the estimation accuracy can be better compared to a method using conventional input scaling and input variable selection methods. The effectiveness of the proposed methods were confirmed through a numerical example and industrial applications in a pharmaceutical process and a distillation process. In the industrial applications, The proposed methods successfully improved the estimation accuracy

in case studies in pharmaceutical and distillation processes.

2.1 Introduction

Although input variable scaling does not affect the result of multiple linear regression (MLR), it affects the estimation result of many other methods such as partial least squares (PLS) and kernel-based methods. Thus, input scaling factor matrix $\mathbf{\Lambda} \in \mathfrak{R}^{M \times M}$ in equations (1.4) and (1.5) should be carefully determined to construct accurate soft-sensors. Though the importance of input variable scaling has been pointed out in literature [4, 63, 64], researches on input variable scaling have not been actively conducted. To our best knowledge, very few papers [65, 66] have proposed input variable scaling methods for soft-sensor design, and conventional methods such as autoscale and range scaling are commonly used [4, 5]. Kuzmanovski et al. [65] used genetic algorithm (GA) to optimize the input scaling factor; however, the computational burden of GA is considerable. Martens et al. [66] proposed to use the magnitude of undesired signals in measurements to determine the input scaling factors. But, their method is applicable only for spectroscopic data. In addition, autoscale and range scaling have a problem as described below. When they are applied, it is assumed that the input variables are equally important for model construction, which is not always the case. In general, some input variables are irrelevant to an output variable, and the others are correlated to the output variable even after input variable selection. This chapter proposes two input variable scaling methods to improve the accuracy of soft-sensors. The proposed methods can determine the input scaling factor λ_m based on the importance of input variables.

The rest of this chapter is organized as follows. In section 2.2, two input variable scaling methods are proposed. Section 2.3 describes an illustrative numerical example. Section 2.4 shows industrial applications in pharmaceutical and distillation processes. Section 2.5 concludes this chapter.

2.2 Method

As mentioned, conventional input scaling methods such as autoscale and range scaling do not determine the input scaling factors based on the importance of individual input variables. More specifically, input variables which have no influence on an output variable are likely to have considerable effect on output estimation because of overfitting especially when the number of samples is small. One can reduce the effect of irrelevant input variables by assigning small input scaling factors to those input variables. On the other hand, large input scaling factors should be assigned to input variables which have a large influence on an output variable.

This research proposes two methods to evaluate the influence of input variables on an output variable and assign the appropriate input scaling factors to input variables. The first one statistically derives the input scaling factors, while the second utilizes spectroscopic data such as absorption spectrum of a material whose content is an estimation target.

2.2.1 Proposed method 1: data-based approach

Proposed method 1 uses regression coefficients of input variables in a statistical model since the regression coefficients are likely to represent the influence of each input variable on an output variable. A detailed algorithm of proposed method 1 is as follows.

1. Prepare an input variable matrix $\boldsymbol{X}$ and an output variable vector $\boldsymbol{y}$.
2. Set an iteration number i to 1 and the maximum iteration number to $I_{\max}$.
3. Set the input scaling factor λ_m as the reciprocal of standard deviation of the mth input variable ($m = 1, 2, \cdots, M$).
4. Derive $\tilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{\Lambda}$.

5. Construct a statistical model, for example a partial least squares (PLS) model, by using the scaled input matrix $\tilde{\boldsymbol{X}}$ and the output variable vector $\boldsymbol{y}$.

6. Finish the calculation if $i = I_{\max}$ or all the input scaling factors converge. Otherwise set i to $i + 1$.

7. Update the input scaling factor λ_m to the product of the regression coefficient of the mth input variable of the statistical model and the standard deviation of the mth input variable in $\tilde{\boldsymbol{X}}$.

8. Go to step 4.

Steps 1-4 correspond to autoscale. Since the regression coefficients depend on the scales of the input variables, the input scaling factors are defined as the products of the regression coefficients and standard deviations of input variables in step 7. Other statistical indexes such as variable influence on projection (VIP) score are also applicable instead of the regression coefficients.

2.2.2 Proposed method 2: knowledge-based approach

In the pharmaceutical and food industries, soft-sensors, whose input variables are spectroscopic data of products, are often used to estimate the content of an important material in the products. In such cases, a spectrum of the important material would also be available. The content of the material is expected to correlate to the (preprocessed) absorbance at the peaks of the spectrum, and not to correlate the (preprocessed) absorbance at the wavelengths where there is no peak. Thus, the input scaling factor λ_m is defined as

$$\lambda_m = \frac{|\xi_m|}{\sigma_{x_m}} \tag{2.1}$$

where ξ_m is the (preprocessed) absorbance of an important material at the mth wavelength and σ_{x_m} is the standard deviation of the absorbance of products at the mth wavelength in

the raw input variable matrix $\boldsymbol{X}$. $|\xi_m|$ is divided by σ_{x_m} so that the standard deviations of absorbance after scaling become $|\xi_m|$.

2.3 Numerical example

In this section, an illustrative numerical example is shown to confirm that input scaling methods can have significant influence on the estimation accuracy of soft-sensors, and proposed method 1 can improve estimation accuracy.

2.3.1 Problem setting

An input-output relationship is defined as follows. The number of input and output variables is 30 and 1, respectively.

$$w_m \sim N(0, 0.005^2) \quad (m = 0, 1, \cdots, 30) \tag{2.2}$$

$$s_m \sim \mathrm{rand}(0, 1) \quad (m = 1, 2, \cdots, 30) \tag{2.3}$$

$$x_m = s_m + w_m \tag{2.4}$$

$$y = s_1 + 3s_2 + 5s_3 + w_0 \tag{2.5}$$

Here, $N(\mu, \sigma^2)$ denotes the normal distribution whose mean is μ and standard deviation is σ, and rand(a, b) denotes the uniform random distribution in a closed interval $[a\ b]$. Thirty samples generated from equations (2.2)-(2.5) are used as model construction data and 3000 samples are used as model validation data. To evaluate the effect of the distribution of the sample on the estimation performance, 1000 sets of model construction and validation data are generated and used separately. PLS is used to construct soft-sensors with the following input scaling methods.

1. Autoscale.

2. A reference method in which $\lambda_m = 1$ $(m = 1, 2, 3)$ and $\lambda_m = 0.1$ $(m = 4, 5, \cdots, 30)$.

3. Proposed method 1 with different maximum iteration numbers I_{max} = 1, 3, 5 and 7.

The number of the latent variables of each PLS model is determined by leave-one-out cross validation. In this example, only three input variables (x_1-x_3) are related to the output variable and the input-output relationship is linear. However, 27 variables (x_4-x_{30}) which are not related to the output variable are used for model construction, and the number of samples for model construction is relatively small, and probability of chance correlation is high. In this example, input variable selection methods were not used to simulate the situation where irrelevant variables are selected by input variable selection methods and to check whether input variable scaling can reduce the risk of chance correlation.

2.3.2 Results and discussion

The model validation result for 1000 sets of model construction and validation data is shown in Figure 2.1. Comparing autoscale and the reference method confirms that estimation accuracy can be greatly improved by properly setting the input scaling factors. In addition, proposed method 1 successfully reduced the mean of root mean square error (RMSE) as well as the reference method. The standard deviations of RMSEs derived by proposed method 1 are larger than that of the reference method because the regression coefficients are derived by only 30 samples and they do not always accurately represent the importance of the input variables. Figure 2.2 shows an example of the change of the regression coefficients for input variables which are not scaled. The values at iteration number 0 are those obtained by autoscale. Figure 2.2 confirms that the regression coefficients converge to the accurate values with 4 or 5 iterations.

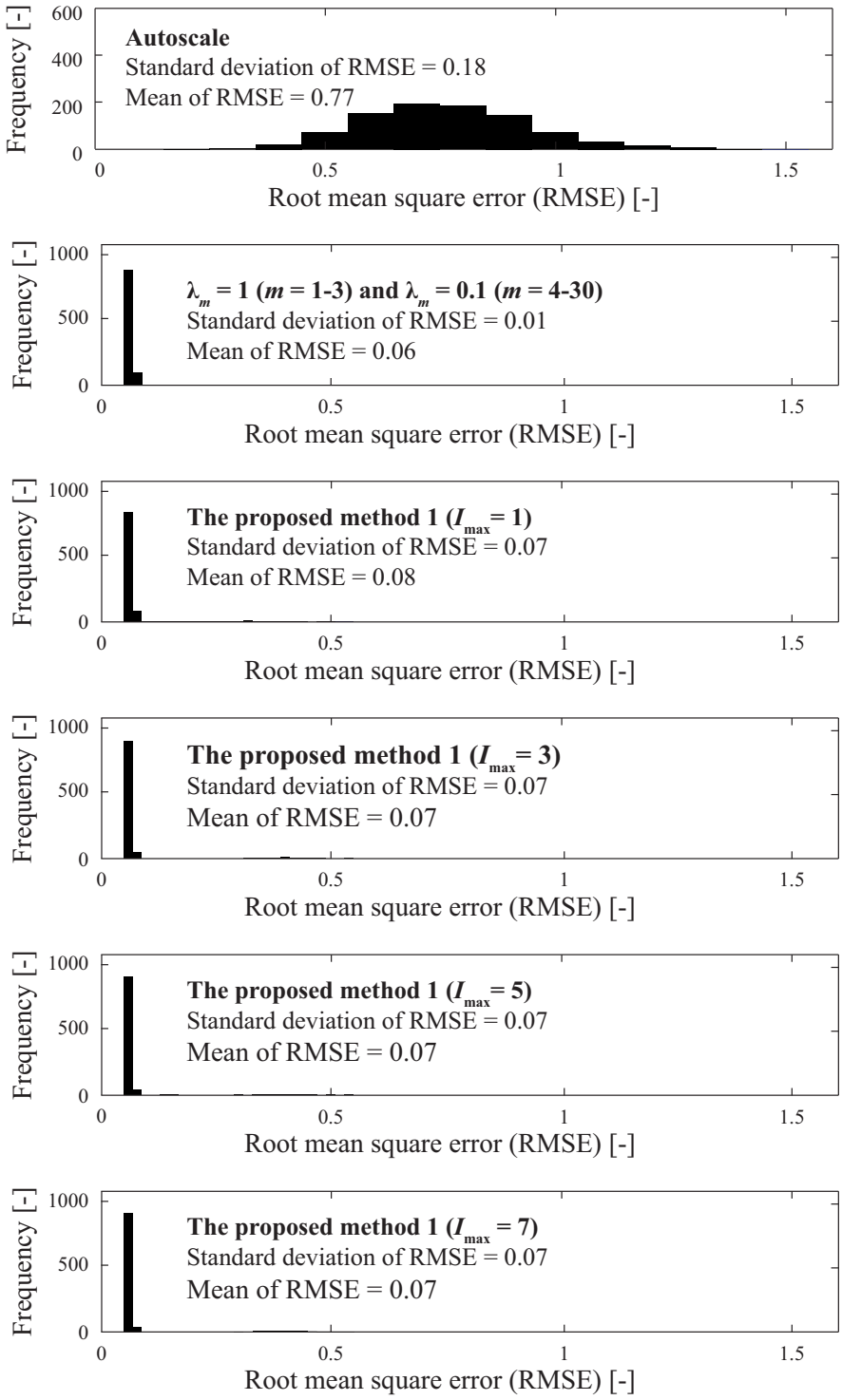

Figure 2.1: Model validation result for 1000 datasets in the numerical example

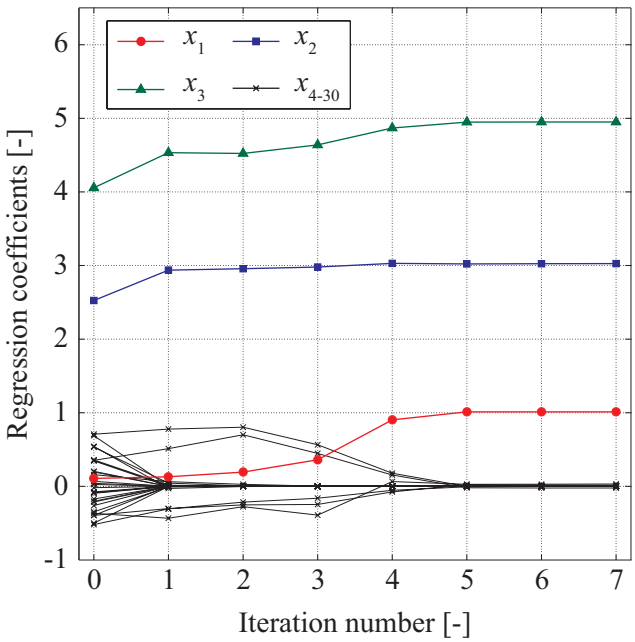

Figure 2.2: Change of regression coefficients for input variables before input scaling with the iteration number

2.4 Application

2.4.1 Pharmaceutical process

In the pharmaceutical industry, the documents on Quality by Design (QbD) and Process Analytical Technology (PAT) [67–70] were published by Food and Drug Administration (FDA) and International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). Since then, online process monitoring and control technologies have attracted much attention. Near infrared spectroscopy (NIRS) is a powerful online monitoring method because of its noninvasiveness and short measuring time; the researches on estimation of many kinds of material attributes such as water content during granulation, blend uniformity, content uniformity and coating thickness by using NIR spectra have been actively conducted [71, 72]. Other spectroscopic data such as Fourier-transform infrared spectroscopy (FT-IR) are used as well.

In this section, soft-sensors were developed to monitor the amount of residual drug substances in manufacturing equipments after cleaning for product quality assurance and safety. More specifically, the amount of magnesium stearate, which is a standard excipient of tablets, was estimated by using absorbance of methanol solutions of different concentrations of magnesium stearate in the infra-red region. The overview of the experimental data is shown in Table 2.1. The absorbance of the solutions were measured from wavelength at 2500-25000 nm. The spectrum of magnesium stearate and methanol solutions of different concentrations of magnesium stearate are shown in Figure 2.3. Here, every spectrum is secondary differentiated and scaled. More detailed information about the materials and experimental condition is described in Nakagawa et al. [61].

Table 2.1: Experimental data for estimation of magnesium stearate amount

Run number	Number of samples	Amount of magnesium stearate [μg/cm^2]
1	15	0.08
2	15	0.20
3	5	0.40
4	5	0.80
5	5	1.20
6	5	1.60
7	15	2.88
8	5	3.20
9	15	4.00
10	15	0.12
11	15	0.24
12	5	0.40
13	5	0.80
14	5	1.20
15	5	1.60
16	15	0.16
17	15	0.32
18	5	0.40
19	5	0.80
20	5	1.20
21	5	1.60

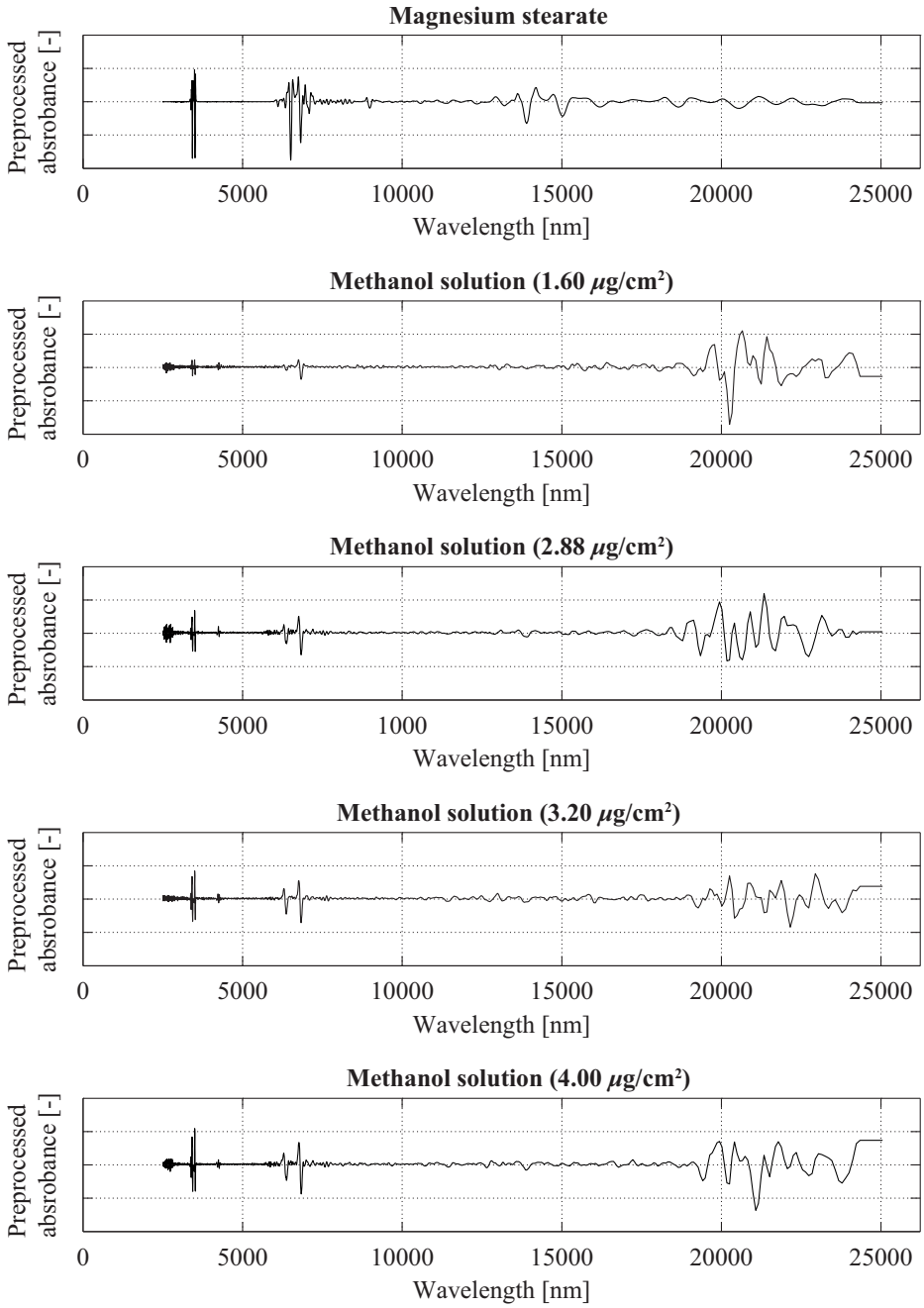

Figure 2.3: Spectra of magnesium stearate and methanol solutions of different concentrations of magnesium stearate

Table 2.2 and Figure 2.4 show the model validation results. Figure 2.5 shows the input scaling factors derived by proposed method 1 in case 3. In this case study, no scaling and autoscale were applied, and all variables (absorbance), those selected on the basis of the peak positions of the spectrum of methanol solutions (absorbance at 3332-3561 nm and 6057-7343 nm), and those selected by two popular statistical input variable selection methods, i.e. variable influence on projection (VIP) [20] and least absolute shrinkage and selection operator (Lasso) [19] were used for model construction. In addition, the data from runs 1-9, 10-15 and 16-21 in Table 2.1 were used for model construction, parameter tuning and model validation, respectively. To evaluate the influence of the number of samples on estimation accuracy, a different number of the model validation samples were used; all samples, 60% of samples, and 20% of samples in each run of model construction data were used in cases 1, 2 and 3, respectively. Tuning parameters such as the numbers of the latent variables in PLS models and thresholds in VIP and Lasso were determined by trial and error so as to minimize RMSE for the parameter tuning data. The proposed methods derived 15-83% smaller RMSE of validation (RMSEV) than the conventional input scaling methods even when input variable selection was conducted by using VIP and Lasso. In addition, RMSEs derived by the proposed methods did not greatly depend on the number of the samples for model construction. The estimation performance derived by the proposed methods were comparable to those derived by using the input variables selected on the basis of peak positions of the spectrum of methanol solutions.

Table 2.2: Results of the case study in the pharmaceutical process.

Method	Scaling	Input variable selection	Model	RMSE (tuning / validation)		
				Case 1 (N = 85)	Case 2 (N = 51)	Case 3 (N = 17)
1	-	-	PLS	0.46 / 0.50	0.82 / 0.84	1.32 / 1.28
2	-	Peak position	PLS	0.21 / 0.22	0.22 / 0.22	0.26 / 0.25
3	-	VIP	PLS	0.45 / 0.48	0.81 / 0.81	1.29 / 1.22
4	-	Lasso	Lasso	0.41 / 0.51	0.72 / 0.78	1.28 / 1.24
5	Autoscale	-	PLS	0.36 / 0.49	0.39 / 0.59	0.46 / 0.64
6	Autoscale	Peak position	PLS	0.24 / 0.26	0.27 / 0.22	0.29 / 0.27
7	Autoscale	VIP	PLS	0.32 / 0.37	0.36 / 0.46	0.46 / 0.62
8	Autoscale	Lasso	Lasso	0.28 / 0.27	0.36 / 0.31	0.38 / 0.34
9	Proposed method 1	-	PLS	0.22 / 0.23	0.23 / 0.22	0.22 / 0.22
10	Proposed method 2	-	PLS	0.20 / 0.26	0.22 / 0.28	0.23 / 0.28

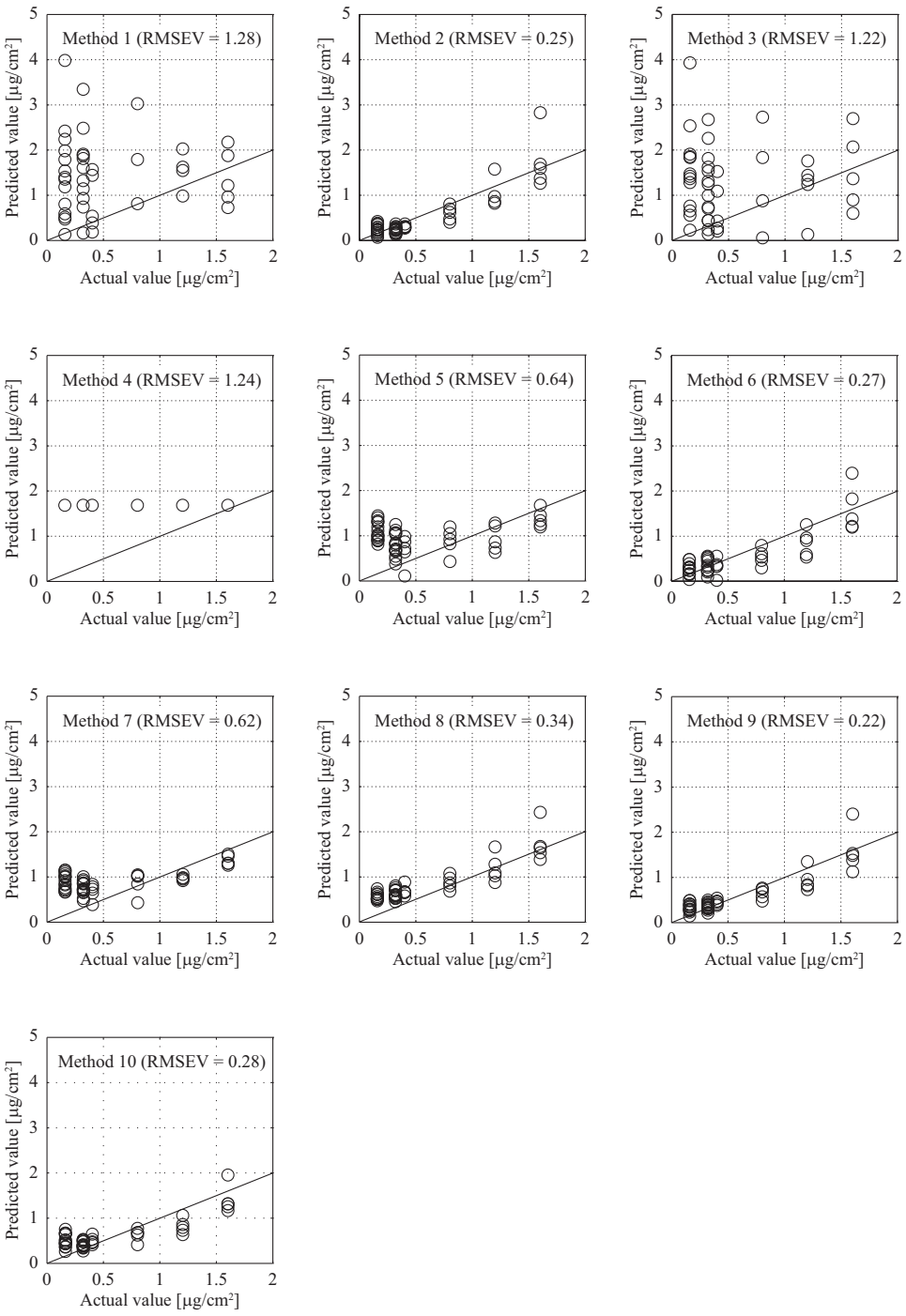

Figure 2.4: Model validation result in the pharmaceutical process (case 3)

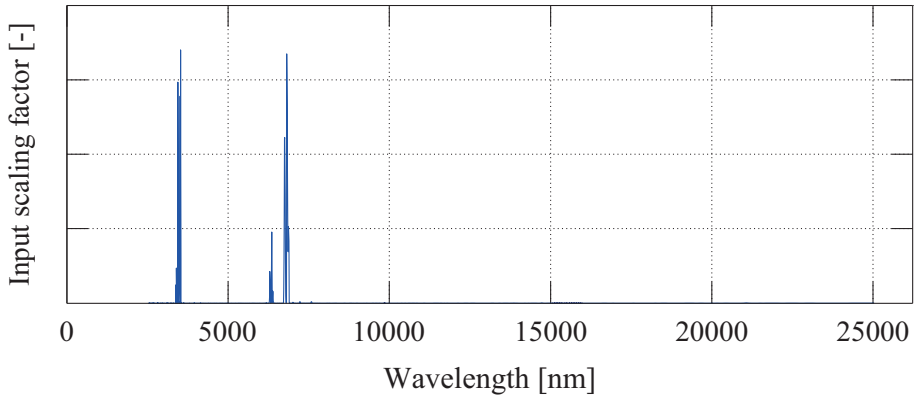

Figure 2.5: Input scaling factors derived by proposed method 1 (case 3)

2.4.2 Distillation process

In distillation processes, soft-sensors are often used to estimate product quality such as the concentration of impurities. In this study, soft-sensors were constructed to estimate 95% distillation temperature, which is an important quality of gasoline. In the target process, 95% distillation temperature is usually measured once a day, and a soft-sensor is needed to realize inferential control of 95% distillation temperature and to reduce the energy consumption. Seventy input variables such as flow rate, temperature and pressure were measured in this process, and 21 variables were removed since they had many abnormal values. The remaining 49 input variables were used for model construction. Two-hundred samples were used as model construction data. Data for parameter tuning and model validation consisted 50 samples, respectively. By using this data, the following methods were compared.

1. Autoscale without input variable selection.
2. Autoscale with VIP.
3. Autoscale with Lasso.
4. Proposed method 1 input variable selection.

Tuning parameters such as the numbers of the latent variables in PLS models and thresholds in VIP and Lasso were selected by trial and error so as to minimize RMSE for the parameter tuning data.

Table 2.3 and Figure 2.6 show the model validation results. The values of 95% distillation temperature were scaled so that RMSEV of proposed method 1 was 1. As shown in Table 2.3 and Figure 2.6, proposed method 1 greatly improved the RMSEV compared to other methods. When autoscale was used, RMSEVs were much larger than RMSE of parameter tuning even though input variable selection was conducted. This means that overfitting occurred because of inappropriate input variable scaling. However, proposed

method 1 derived small RMSEs both for parameter tuning and validation data by adjusting the input scaling factors based on the importance of individual input variables. This result confirmed the usefulness of proposed method 1 as well as the previous case studies.

Table 2.3: Results of the case study in the distillation process

Method	Scaling	Input variable selection	Model	RMSE (tuning / validation)
1	Autoscale	-	PLS	1.15 / 1.84
2	Autoscale	VIP	PLS	1.15 / 1.84
3	Autoscale	Lasso	Lasso	1.18 / 1.45
4	Proposed method 1	-	PLS	1.08 / 1.00

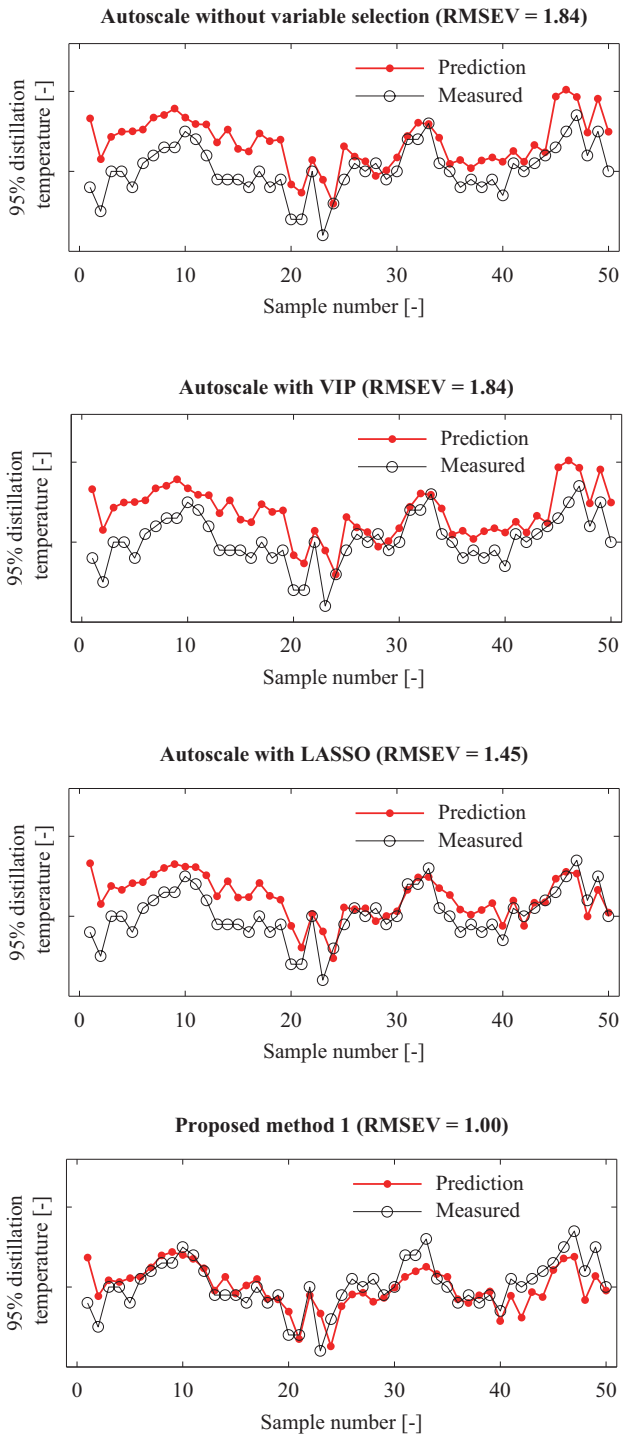

Figure 2.6: Model validation result in the distillation process

2.5 Conclusions

Input variable scaling methods for soft-sensor design were investigated in this chapter, and it was revealed that the input scaling factors should be determined on the basis of the importance of input variables. In addition, two novel input scaling methods, which can evaluate the importance of input variables, were proposed. One method statistically derives the input scaling factors. The other one utilizes spectroscopic data of a material whose content is an estimation target. The effectiveness of the proposed methods was confirmed through its application to a numerical example and industrial applications in pharmaceutical and distillation processes.

Chapter 3

Input Variable Selection for Batch Processes

Abstract

Although numerous input variable selection methods have been proposed, trial and error is unavoidable in practice because it is not clearly understood how to choose the best input variable selection method. One of the reason for this situation is that most of past input variable selection methods do not take into account properties of target processes. In this chapter, an input variable selection method for batch processes is proposed. The proposed method evaluates input variables based on their standard deviations in the same batch and different batches. The proposed input variable selection method was applied to a real pharmaceutical process data and the estimation accuracy was improved by 28.7 and 45.5% in root mean square error of validation (RMSEV) compared to variable influence on projection (VIP) and a method based on engineers' process knowledge, respectively.

3.1 Introduction

To construct an accurate soft-sensor, it is crucial to select an appropriate set of input variables. However, it is quite difficult when the number of measured variables are large and a target process is not well understood from the chemical and physical view point. To efficiently select input variables, many methods have been proposed as explained in section 1.2.4. However, input variable selection is still one of the time-consuming steps of soft-sensor design because it is not clear which method is the suitable to individual cases. This might be because most of past input variable selection methods are based only on statistical property of the data and do not take into account properties of target processes, and thus the interpretation and the improvement of the input variable selection results are difficult. To solve this problem, an input variable selection method for batch processes, which are important for the production of high value-added products such as fine chemicals and drugs, is proposed in this chapter. The proposed method takes into account the property of batch processes, thus it can improve the estimation accuracy and is easy to understand.

The rest of this chapter is organized as follows. In section 3.2, the proposed input variable selection method is explained. In section 3.3, the performance of the proposed method is evaluated through applying it to a blending process. Finally, conclusions are described in section 3.4.

3.2 Method

In batch processes, different products can be produced in different batches and multiple samples can be taken at the same time in a batch. The samples taken at the same time are ideally the same, thus, the difference between each input variable in the samples can be regarded as undesirable fluctuation of the input variable. Hence, input variables with

large standard deviation are not useful for output estimation. On the other hand, each input variable measured at the same processing time should be different in batches with different outputs since output estimation becomes the same if input variable is the same. Thus, input variables with large standard deviation between batches are useful for output estimation.

From the above discussion, the following method is developed. In the following method, it is assumed that sampling is conducted once during each batch, and the sampling time in every batch is the same.

1. Calculate mean and variance of the mth input variable in the kth batch.

$$\bar{x}_{*mk} = \frac{1}{N_k}\sum_{n=1}^{N_k} x_{nmk} \tag{3.1}$$

$$\mathrm{V}_n(x_{nmk}) = \frac{1}{N_k-1}\sum_{n=1}^{N_k}(x_{nmk}-\bar{x}_{*mk})^2 \tag{3.2}$$

Here, the nth measurement of the mth input variable in the kth batch is denoted by x_{nmk} ($n = 1,2,\cdots,N_k$, $m = 1,2,\cdots,M$, and $k = 1,2,\cdots,K$), where N_k and K denote the number of samples in the kth batch and the number of batches, respectively.

2. Select the input variables which satisfy the following condition.

$$\eta_m = \frac{\mathrm{V}_k(\bar{x}_{*mk})}{\sum_{k=1}^{K}\mathrm{V}_n(x_{nmk})} > \alpha \tag{3.3}$$

$$\mathrm{V}_k(\bar{x}_{*mk}) = \frac{1}{K-1}\sum_{k=1}^{K}(\bar{x}_{*mk}-\bar{x}_{*m*})^2 \tag{3.4}$$

$$\bar{x}_{*m*} = \frac{1}{K}\sum_{k=1}^{K}\bar{x}_{*mk} \tag{3.5}$$

where α denotes a threshold for input variable selection.

3.3 Application to blending process

This section shows an application example of the proposed method to a blending process of a pharmaceutical process. The estimation objective is active pharmaceutical ingredient (API) content in granules for tableting, which is generally not measured. The input variable is absorbance of the granules at near-infrared (NIR) region (800-2500 nm). If API content in granules can be estimated by using a soft-sensor, the operation condition of the following processes can be changed to make API content in the final products satisfy the specification.

3.3.1 Experimental

Figure 3.1 shows the overview of the experimental procedure. Eighteen blending experiments, in which six powders were blended, were conducted with different API content using a 3 L scale V-blender (Tsutsui Scientific Instruments Co., Ltd.). The amount of the other materials were constant. After each blending experiment, the granules for tableting were taken out and 200 mg of the granules were set in vials, absorbance (2203 points) was measured with MPA (Bruker Optics K. K.), and API content was measured with Alliance Waters 2690 Separations Module (Waters Corporation). The overview of the experimental data is shown in Table 3.1. In addition, Figure 3.2 shows the spectra of each component and a granule for tableting. In this study, the data of batches 1-8 are the model construction data, the data of batches 9-16 are the parameter tuning data, and the data of batches 17 and 18 are the model validation data.

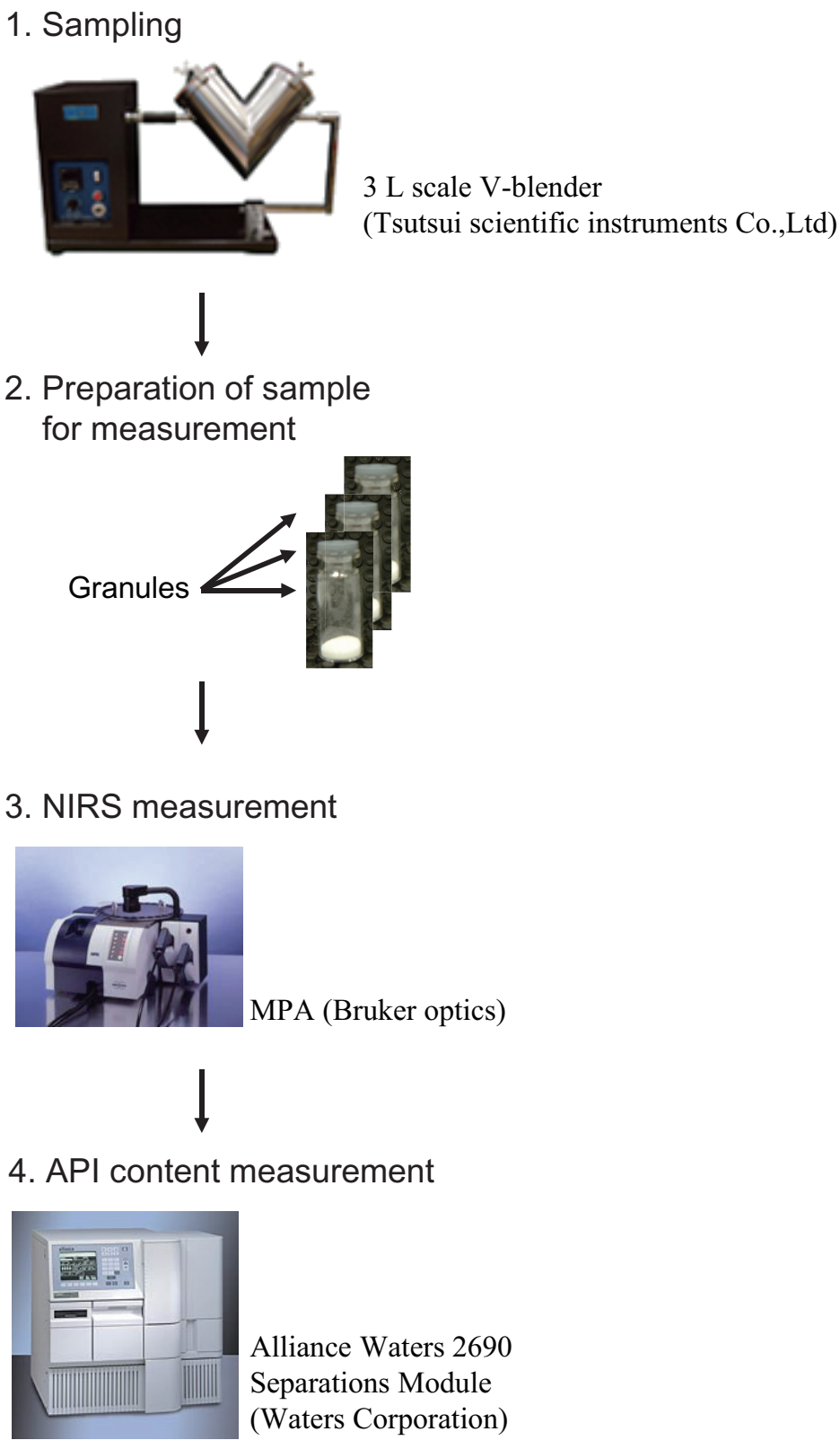

Figure 3.1: Procedure of blending experiments

Table 3.1: Experimental data

Batch	Number of samples	Mean of API content [%]
1	90	68.1
2	86	83.0
3	100	88.7
4	20	97.4
5	10	98.6
6	90	107.7
7	90	113.8
8	90	128.3
9	10	73.9
10	10	94.0
11	10	96.8
12	10	98.3
13	10	98.8
14	10	99.5
15	10	100.1
16	10	122.9
17	10	96.0
18	10	100.0

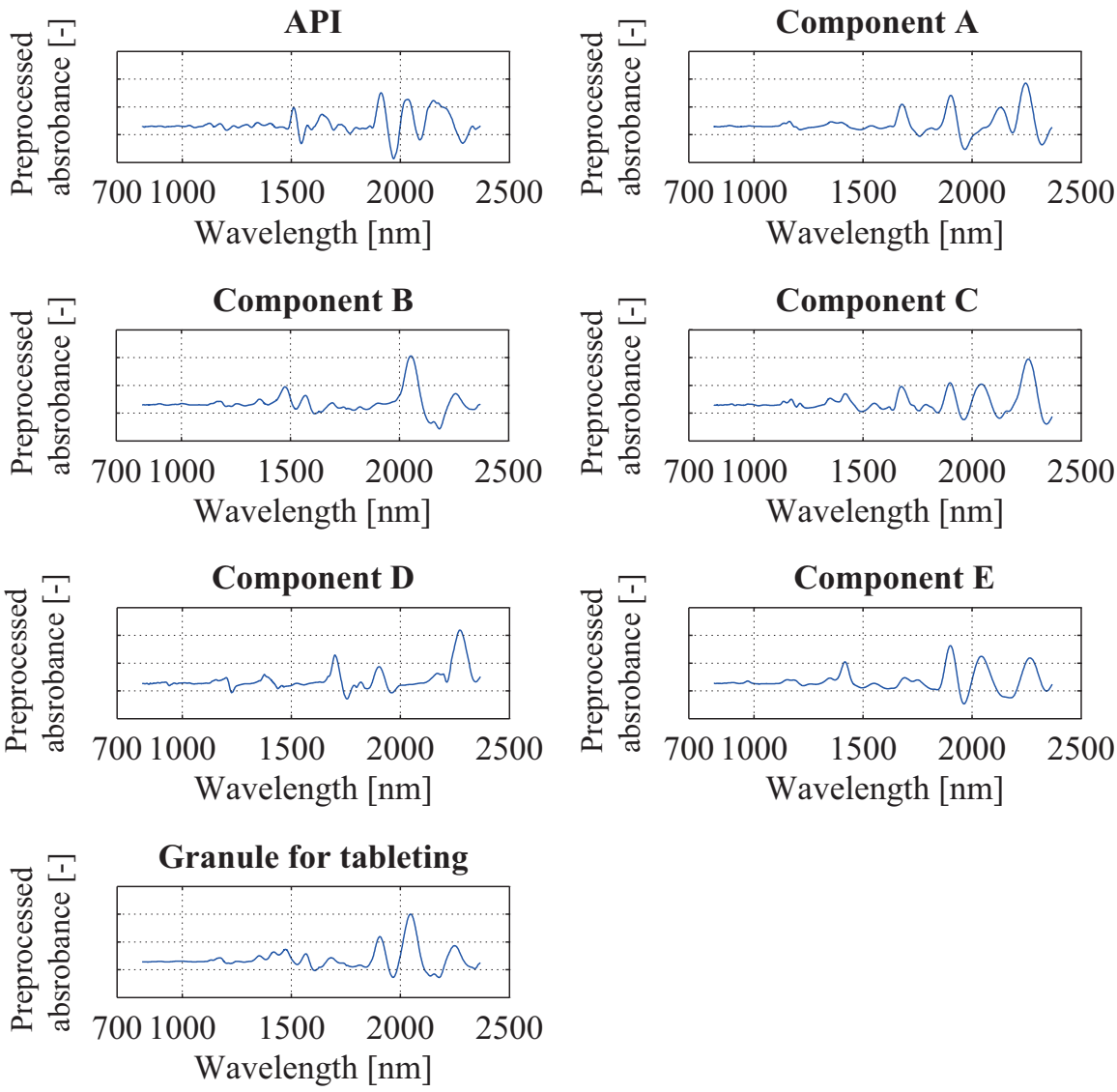

Figure 3.2: Spectra of a granule for tableting and its components

3.3.2 Data analysis

The detailed comparison procedure of variable selection methods is as follows.

1. Preprocessing

   Apply first order differential using Savitsky-Golay filter [73] and standard normal variate (SNV) to NIR spectra data. By using Savitsky-Golay filter, the effect of the noise on NIR spectra can be reduced. SNV can correct the variance in light path length caused by changes in the particle size and density [74]. In this application, the window size and the polynomial order in Savitsky-Golay filter were 117 and 5, respectively.

2. Input variable (wavelength) selection

   Use the input variables selected by the proposed method, engineers in a pharmaceutical company based on process knowledge and trial and error, and variable influence on projection (VIP), a widely used variable selection method. The engineers selected 1081-1132 nm, 1216-1249 nm and 1263-1301 nm.

3. Model construction

   Construct estimation models by using locally weighted PLS (LW-PLS) in which the following similarity is used.

$$\omega_n = \exp\left(-\frac{d_n}{\sigma_d \varphi}\right) \tag{3.6}$$

$$d_n^2 = (\boldsymbol{x}_n - \boldsymbol{x}_q)^{\mathrm{T}}(\boldsymbol{x}_n - \boldsymbol{x}_q) \tag{3.7}$$

   Here, φ is a tuning parameter, σ_d is a standard deviation of d_n ($n = 1, 2, \cdots, N$), $\boldsymbol{x}_n$ is the nth sample, and $\boldsymbol{x}_q$ is a query, for which an output estimation is required. LW-PLS constructs local PLS models by prioritizing the similar samples. The detailed algorithm of LW-PLS is explained in section 4.2.

Model parameters in each model, i.e. the localization parameter φ, the threshold for the proposed variable selection index α_1, the threshold for the VIP score α_2, and the number of latent variables R, were determined so that root mean square error for parameter tuning data becomes the minimum.

3.3.3 Results and discussion

Table 3.2 shows the selected parameters, root mean square error of parameter tuning (RMSET) and root mean square error of validation (RMSEV). In addition, model validation results are shown in Figure 3.3. By using the proposed wavelength selection method, RMSEV was improved by 28.7 and 45.5% compared to VIP and the engineers' method, respectively. The results of the case study demonstrate the usefulness of the proposed input variable selection method.

Since the spectra of the granules were dealt with in this case study, Lambert-Beer law is not satisfied and input-output relationship is complicated. This might be the main reason for the bad estimation performance of the engineers' method.

Figure 3.4 (top) shows the VIP score (R = 11) and preprocessed API spectrum. VIP score mostly has a correlation with absorbance values of API spectrum and it can be expected that RMSE becomes small when the threshold for VIP α_2 is large. However, the minimum RMSET was obtained by using all wavelengths when VIP was applied. This result implies that VIP dose not properly evaluate the importance of the input variables.

On the other hand, the proposed method selected 259 wavelengths, which had index η larger than 10. In addition, the index η and preprocessed API spectrum are shown in Figure 3.4 (middle), and $\mathrm{V}_k(\bar{x}_{*mk})$ and $\sum_{k=1}^{K} \mathrm{V}_n(x_{nmk})$ in equation (3.3) are shown in Figure 3.4 (bottom). η does not have a strong correlation with absorbance values of API spectrum because η takes account not only of the effect of API content on NIR spectra but also of the effect of other factors on NIR spectra. The wavelengths around 1910 and

Table 3.2: Results of the case stuy in the blending process

Method	♯	α_1	α_2	φ	R	RMSET	RMSEV
Engineers' method	227	-	-	0.8	1	2.49	2.24
VIP	2087	-	0	10	11	2.14	1.71
Proposed method	259	10	-	0.5	9	1.96	1.22

♯: the number of selected wavelengths

1970 nm were not selected although peak absorbance values of API and $\mathrm{V}_k(\bar{x}_{*mk})$, which is the effect of API content on NIR spectra, were large. This is because $\sum_{k=1}^{K} \mathrm{V}_n(x_{nmk})$, which is the effect of other factors on NIR spectra, was also large at these wavelengths. On the contrary, the wavelengths around 1120 and 1190 nm were selected although peak absorbance values of API and $\mathrm{V}_k(\bar{x}_{*mk})$ were small.

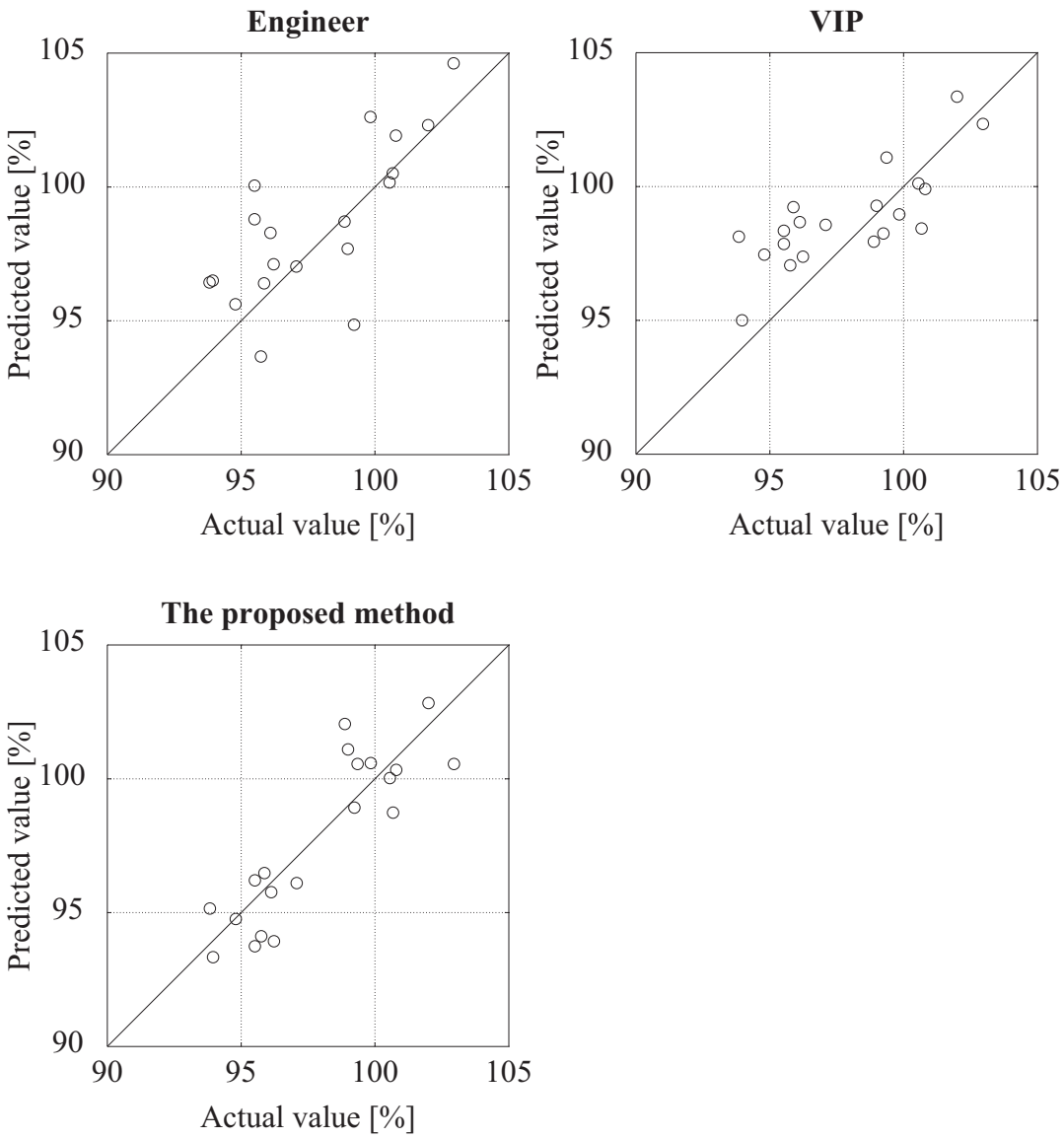

Figure 3.3: Model validation results in the blending process

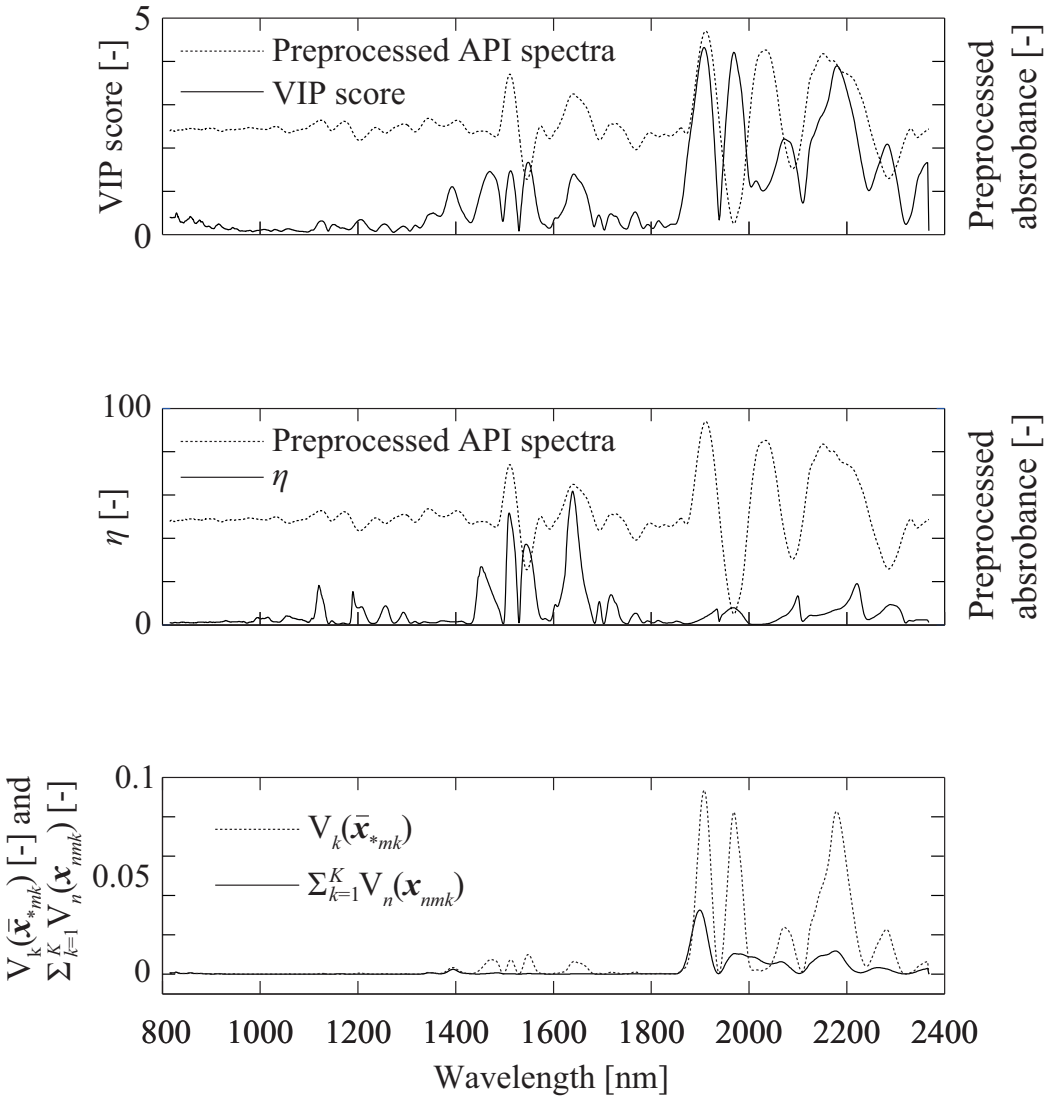

Figure 3.4: VIP score and preprocessed API spectra (top), input variable selection index η and preprocessed API spectra (middle), and $V_k(\bar{x}_{*mk})$ and $\sum_{k=1}^{K} V_n(x_{nmk})$ (bottom)

3.4 Conclusions

In this chapter, an input variable selection method for batch processes is proposed. The proposed method takes into account the property of batch processes, thus estimation accuracy can be improved. The usefulness of the proposed method was confirmed through its application to a blending process in a pharmaceutical process. The estimation error was reduced by 28.7 and 45.5% in root mean square error of validation (RMSEV) compared to variable influence on projection (VIP) and a method based on engineers' process knowledge, respectively. The results clearly show that the proposed method is superior to the conventional ones.

Chapter 4

Application of Just-In-Time Model and Proposal of New Similarity Measure

Abstract

Just-in-time (JIT) models have attracted much attention since they can prevent the performance deterioration of soft-sensors caused by changes in process characteristics and operation condition. In this chapter, successful applications of JIT models integrated with commercial model predictive control (MPC) software in chemical processes are shown. The developed system has adopted locally weighted partial least squares (LW-PLS) to build soft-sensors. LW-PLS is a kind of JIT modeling method that can cope with changes in process characteristics as well as process nonlinearity. Thus, LW-PLS helps engineers to reduce their burden of model maintenance, which has been recognized as the most serious problem in practice. Inferential control systems have been used for more than two years at Showa Denko K.K. (SDK) in Japan, and the operation cost and environmental burden have been significantly reduced. In the cracked gasoline fractionator, for example, about 0.6% of operation cost was cut successfully. Moreover, a novel definition of sim-

ilarity between samples is proposed to improve the estimation performance of LW-PLS. The usefulness of the similarity was confirmed in numerical examples and an industrial application in a distillation process.

4.1 Introduction

Soft-sensors have been already implemented in many industrial processes. To implement more soft-sensors, it is crucial to reduce the burden of maintenance of soft-sensors, which is caused by changes in process characteristics and operation condition. To realize maintenance free soft-sensors, just-in-time (JIT) modeling methods have been investigated. However, few papers have reported long-term application results of process control using JIT soft-sensors in real processes, though it is common practice to evaluate the estimation performance of soft-sensors by using industrial process data. In this chapter, long-term successful applications of locally weighted partial least squares (LW-PLS), a kind of JIT models, are presented. In addition, a new similarity measure based on the weighted Euclidean distance is proposed to enhance estimation accuracy of LW-PLS since it is crucial to define the similarity between samples to construct an accurate soft-sensor.

The rest of this chapter is organized as follows. Section 4.2 explains the algorithm of LW-PLS. The details of the control systems and application results are shown in section 4.3. Section 4.4 proposes a new similarity to further improve the estimation performance of LW-PLS. Sections 4.5 and 4.6 show the application results of the proposed similarity. Finally, this chapter is concluded in section 4.7.

4.2 Locally weighted partial least squares

LW-PLS [53, 75] is a JIT modeling method, which does not construct a regression model off-line. Instead, an input variable matrix $\boldsymbol{X}$ and an output variable vector $\boldsymbol{y}$ are stored in a

database. When an output estimation is required for a query $\boldsymbol{x}_q$, the similarity ω_n between $\boldsymbol{x}_q$ and $\boldsymbol{x}_n$ is calculated, and a local PLS model is constructed by weighting samples with a similarity matrix $\boldsymbol{\Omega}$ defined by

$$\boldsymbol{\Omega} = \mathrm{diag}(\omega_1, \omega_2, \cdots, \omega_N) . \tag{4.1}$$

In general, ω_n is defined on the basis of the distance between samples in the input space.

The output estimate $\hat{y}_q$ is calculated through the following procedure.

1. Determine the number of latent variables R and set r to 1.

2. Calculate the similarity matrix $\boldsymbol{\Omega}$.

3. Calculate $\boldsymbol{X}_r$, $\boldsymbol{Y}_r$ and $\boldsymbol{x}_{q,r}$

$$\boldsymbol{X}_r = \boldsymbol{X} - \boldsymbol{1}_N[\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_M] \tag{4.2}$$

$$\boldsymbol{y}_r = \boldsymbol{y} - \boldsymbol{1}_N \bar{y} \tag{4.3}$$

$$\boldsymbol{x}_{q,r} = \boldsymbol{x}_q - [\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_M]^{\mathrm{T}} \tag{4.4}$$

$$\bar{x}_m = \sum_{n=1}^{N} \omega_n x_{nm} / \sum_{n=1}^{N} \omega_n \tag{4.5}$$

$$\bar{y} = \sum_{n=1}^{N} \omega_n y_n / \sum_{n=1}^{N} \omega_n \tag{4.6}$$

   where $\boldsymbol{1}_N \in \Re^N$ is a vector of ones.

4. Derive the r-th latent variable of $\boldsymbol{X}$

$$\boldsymbol{t}_r = \boldsymbol{X}_r \boldsymbol{w}_r \tag{4.7}$$

   where $\boldsymbol{w}_r$ is the eigenvector of $\boldsymbol{X}_r^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{y}_r \boldsymbol{y}_r^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{X}_r$ which corresponds to the maximum eigen value.

5. Derive the r-th loading vector of $\boldsymbol{X}$

$$\boldsymbol{p}_r = \frac{\boldsymbol{X}_r^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{t}_r}{\boldsymbol{t}_r^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{t}_r} \tag{4.8}$$

and the regression coefficient

$$q_r = \frac{\boldsymbol{y}_r^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{t}_r}{\boldsymbol{t}_r^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{t}_r} \ . \tag{4.9}$$

6. Derive the r-th latent variable of $\boldsymbol{x}_q$

$$t_{q,r} = \boldsymbol{x}_{q,r}^{\mathrm{T}} \boldsymbol{w}_r \ . \tag{4.10}$$

7. If $r = R$, calculate the output estimate

$$\hat{y}_q = \bar{y} + \sum_{r=1}^{R} t_{q,r} q_r \tag{4.11}$$

and finish estimation. Otherwise, set

$$\boldsymbol{X}_{r+1} = \boldsymbol{X}_r - \boldsymbol{t}_r \boldsymbol{p}_r^{\mathrm{T}} \tag{4.12}$$

$$\boldsymbol{y}_{r+1} = \boldsymbol{y}_r - \boldsymbol{t}_r q_r \tag{4.13}$$

$$\boldsymbol{x}_{q,r+1} = \boldsymbol{x}_{q,r} - t_{q,r} \boldsymbol{p}_r \ . \tag{4.14}$$

8. Set r to $r + 1$ and go to step 4.

This algorithm is applicable for multiple outputs just by adding the columns to $\boldsymbol{y}$. When $\boldsymbol{\Omega}$ is an identity matrix, LW-PLS becomes the same as PLS. At step 3, the weighted mean of each variable is subtracted from each column of $\boldsymbol{X}$, $\boldsymbol{y}$ and $\boldsymbol{x}_q^{\mathrm{T}}$ to make the query near to the origin of the multidimensional space. At steps 4-8, the latent variable $\boldsymbol{t}$, the loading vector $\boldsymbol{p}$ and the regression coefficient q are derived iteratively, and the output estimate $\hat{y}_q$ is calculated when $r = R$.

4.3 Industrial applications of locally weighted partial least squares

At Showa Denko K.K. (SDK) in Japan, soft-sensors based on LW-PLS have been applied to various processes. In the early stage of soft-sensor implementation, multiple linear regression (MLR) and PLS had been mainly used. However, the company had to confront two major problems: first, the considerable cost and time required to implement soft-sensors, and second, the burden of the maintaining soft-sensors to prevent the performance deterioration. To solve these problems and increase the number of soft-sensor applications, a practical configuration of an inferential control system was developed by integrating a commercial model predictive control (MPC) software and LW-PLS-based soft-sensors. The developed system can be easily implemented to processes using common software and personal computers as shown in the next section. In addition, by using LW-PLS-based soft-sensors, the burden of the model maintenance can be reduced since they can cope with changes in process characteristics as well as process nonlinearity.

The following sections explain the details of the configuration of the developed inferential control system and two successful results of applying the system to chemical processes. In these applications, the following form of the similarity ω_n is used for the sake of simplicity.

$$\omega_n = \exp\left(-\frac{d_n}{\sigma_d \varphi}\right) \tag{4.15}$$

$$d_n^2 = (\boldsymbol{x}_n - \boldsymbol{x}_q)^{\mathrm{T}}(\boldsymbol{x}_n - \boldsymbol{x}_q) \tag{4.16}$$

Here, σ_d is a standard deviation of d_n $(n = 1, 2, \cdots, N)$ and φ is a localization parameter. The similarity decreases steeply when φ is small and gradually when φ is large. To cope with nonlinearity between input and output variables φ needs to be small, however, LW-PLS models can be sensitive to noise when φ is too small.

4.3.1 Configuration of the inferential control system

Figure 4.1 shows the inferential control system configuration developed at the SDK Oita plant, in which soft-sensors, a commercial MPC software and a distributed control system (DCS) are combined with each other. DCS accumulates measurements of process variables and returns the control signals to the process. The personal computer 1 (PC 1) for MPC receives the information of process variables including controlled variables (CVs), manipulated variables (MVs) and disturbance variables (DVs) from DCS, and returns the optimized values of MVs to DCS. In addition, PC 1 transfers values of input variables of the soft-sensors from the database to the Excel® platform in PC 2. The soft-sensor programs in PC 2 calculate the output estimates using the data in the Excel® plat-

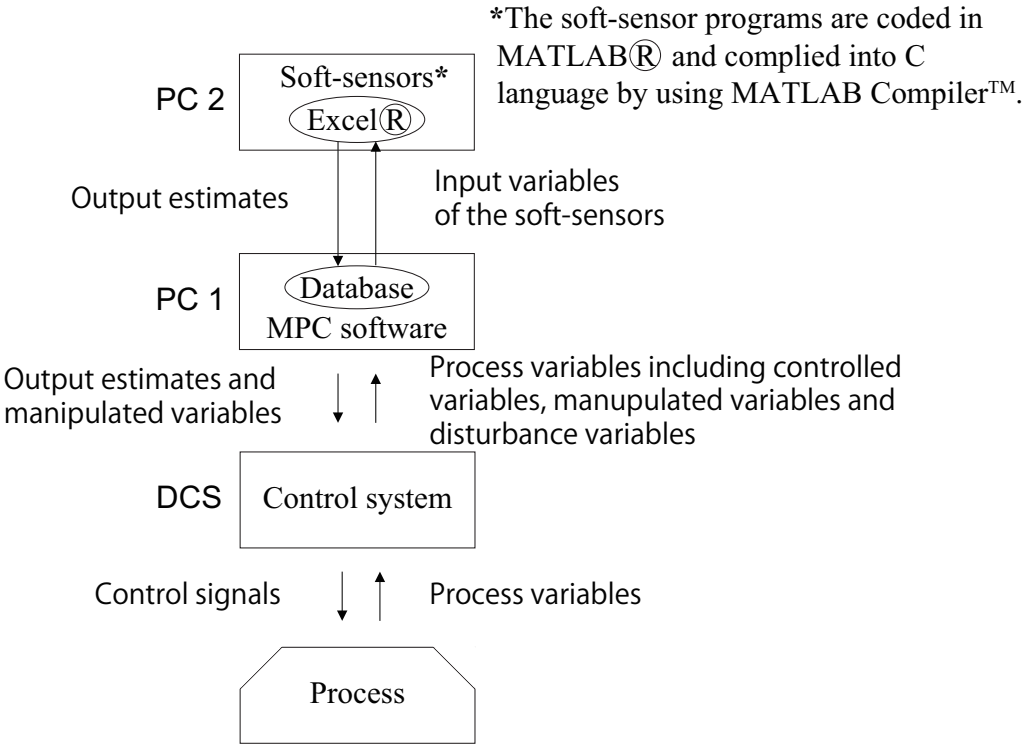

Figure 4.1: Configuration of the developed inferential control system in the SDK Oita plant

form, and the output estimates are returned to PC 1 and DCS. The soft-sensor programs are coded in MATLAB® and compiled into C language by using MATLAB Compiler™ in order to make them available without MATLAB®. The soft-sensor programs and the MPC software can be easily installed in commercial PCs, which can be connected to DCS. In addition, this inferential control system can be applied to any processes. Thus, the developed system can reduce cost and time for implementation.

4.3.2 CGL fractionator of ethylene production process

A schematic diagram of the ethylene production process at the SDK Oita plant is shown in Figure 4.2. Raw materials are fed into the cracking furnace, and many products are produced by using the following towers. In addition, a schematic diagram of the cracked gasoline (CGL) fractionator of the ethylene production process at the SDK Oita plant is shown in Figure 4.3. The concentration of aromatics in the product CGL, which is an important product quality, is usually analyzed only once a day in a laboratory. The concentration has a soft lower bound constraint and should be kept close to the constraint in order to satisfy the product specification and to reduce operation cost. Equivalently speaking, short-time violations are acceptable. The coil outlet temperature (COT) of the cracking furnace is the main manipulated variable for the aromatics concentration control; the aromatics concentration can be increased by increasing COT. However, higher COT requires higher operation cost. Although a PLS-based soft-sensor and MPC had been implemented to estimate and control the aromatics concentration, the estimation performance was not high enough. Thus, COT had been kept excessively higher than the optimal to satisfy the constraint on the aromatics concentration. LW-PLS replaced conventional PLS to enhance the estimation accuracy, decrease COT and lower the operation cost, which is dominated by the energy consumption amount in the cracking furnace.

A commercial MPC software package has been implemented in this ethylene pro-

duction process. The package has various functions for different scales of processes and different purposes. One of the functions of the package is to solve an optimization problem by using a model of the whole ethylene production process. The constraint on the aromatics concentration is taken into account in the optimization problem whereas no MPC controller is implemented in the CGL fractionator itself. Other MPC controllers implemented in other processes in the ethylene production process control each process so that the constraint on the aromatics concentration is satisfied.

Eight process variables shown in Figure 4.3 were selected as input variables of the soft-sensor on the basis of engineers' process knowledge. In addition, COT measured four hours before was used as an input variable together with the selected input variables, since it takes about four hours for materials to reach the CGL fractionator from the cracking furnace. Hence, the total number of input variables is nine. The number of latent variables R and the localization parameter φ were determined by cross validation; R and φ were set to four and 0.5, respectively. When PLS was applied to the aromatics concentration estimation, bias update had been used to reduce estimation error. After LW-PLS was implemented instead of PLS, the 30 newest samples were stored in the database to cope with recent changes in process characteristics. In addition, 395 samples obtained from 1st June 2010 to 31st May 2011 were selected as core data, which had been always stored in the database to prevent overfitting and cope with nonlinearity and abrupt changes in process characteristics. The use of core data is significantly useful to make JIT soft-sensors robust. The number of newest samples and the core data were selected by trial and error.

Figure 4.4 shows the laboratory measurements of the aromatics concentration y_{aroma}, the estimates, the soft constraint and COT, which is the main MV of MPC. All variables are scaled in this figure, and the control results before and after the implementation of LW-PLS are compared.

In addition, Table 4.1 summarizes root mean square errors (RMSE) of prediction,

mean absolute deviation (MAD) of y_{aroma} from the lower bound, standard deviation σ of y_{aroma} and mean of COT. Here, RMSE, $\text{MAD}_{y_{\text{aroma}}}$, $\sigma_{y_{\text{aroma}}}$ and COT are scaled so that they are 100 when PLS was used, i.e. before the implementation of LW-PLS. The unit of COT is Celsius.

All indexes in Table 4.1 were improved by using LW-PLS. About 0.6% of operation cost was reduced by using the MPC system combined with the LW-PLS-based soft-sensor. Although some of the laboratory measurements of aromatics concentration are under the lower bound when LW-PLS is applied, this is acceptable since the violation did not continue for a long time. The control system has been stably working for more than two year although only two months of the data are shown in Figure 4.4.

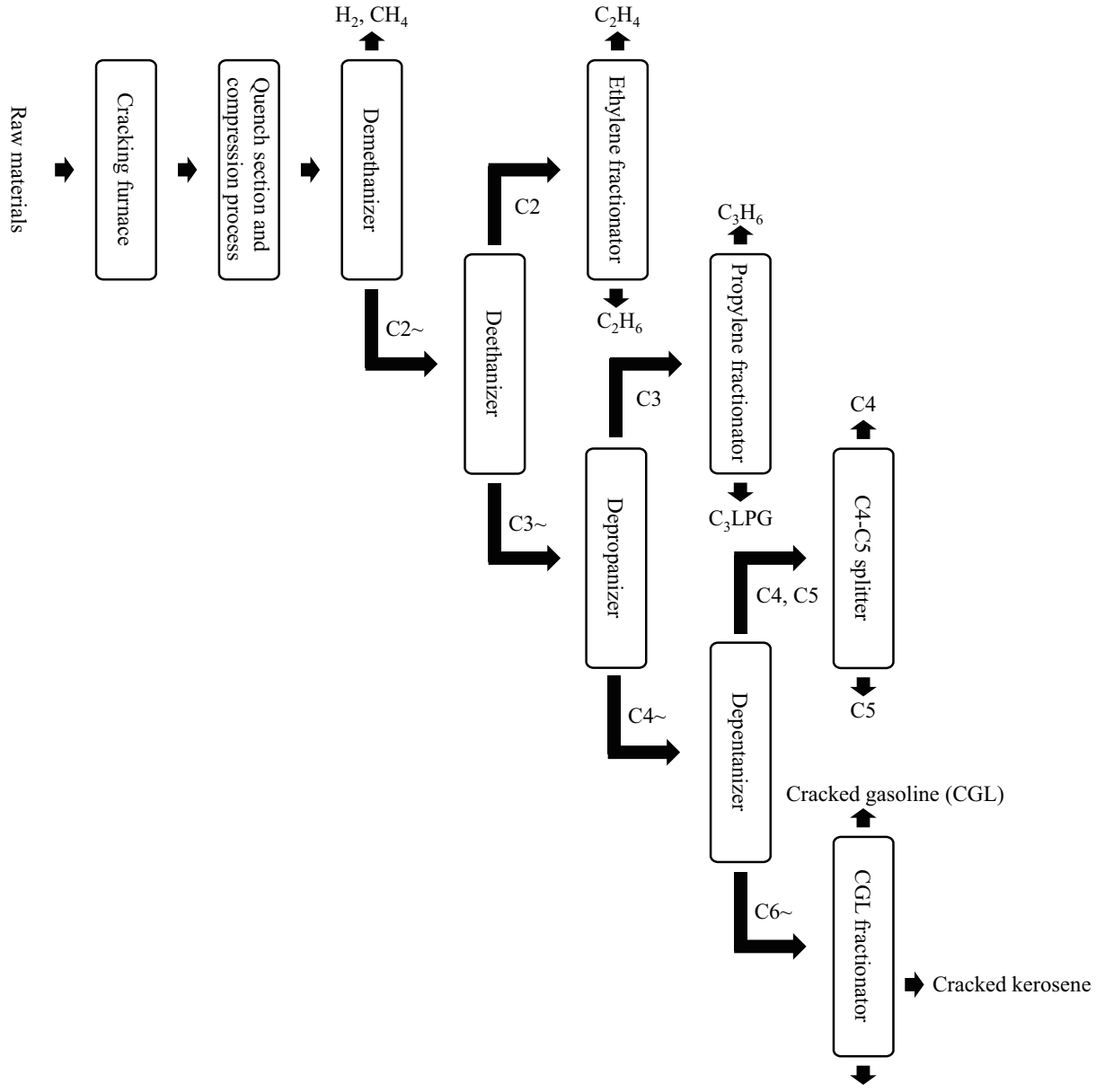

Figure 4.2: Schematic diagram of the ethylene production process at the SDK Oita plant

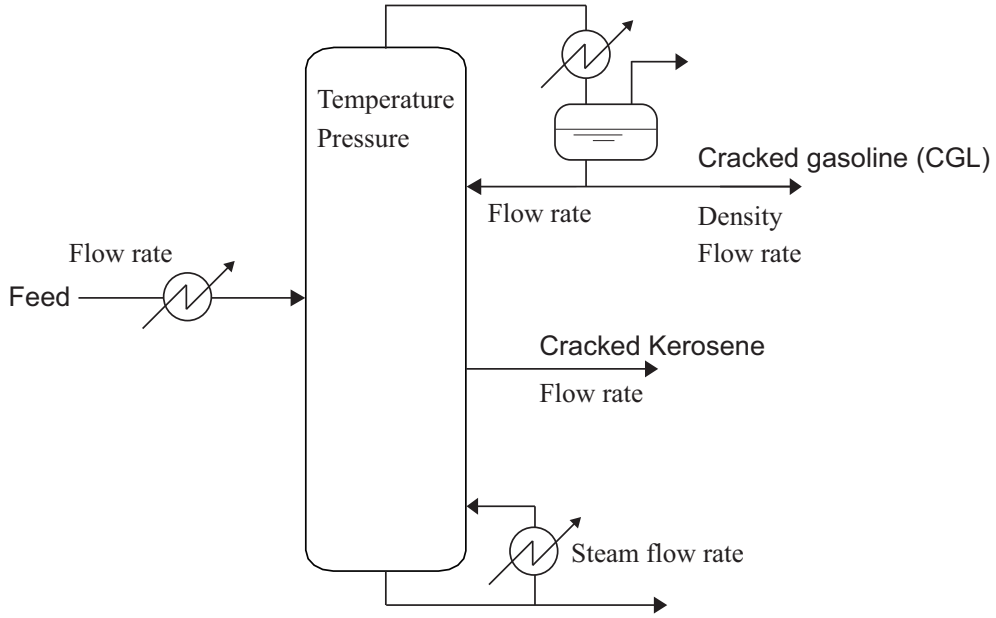

Figure 4.3: Schematic diagram of the CGL fractionator of the ethylene production process at the SDK Oita plant

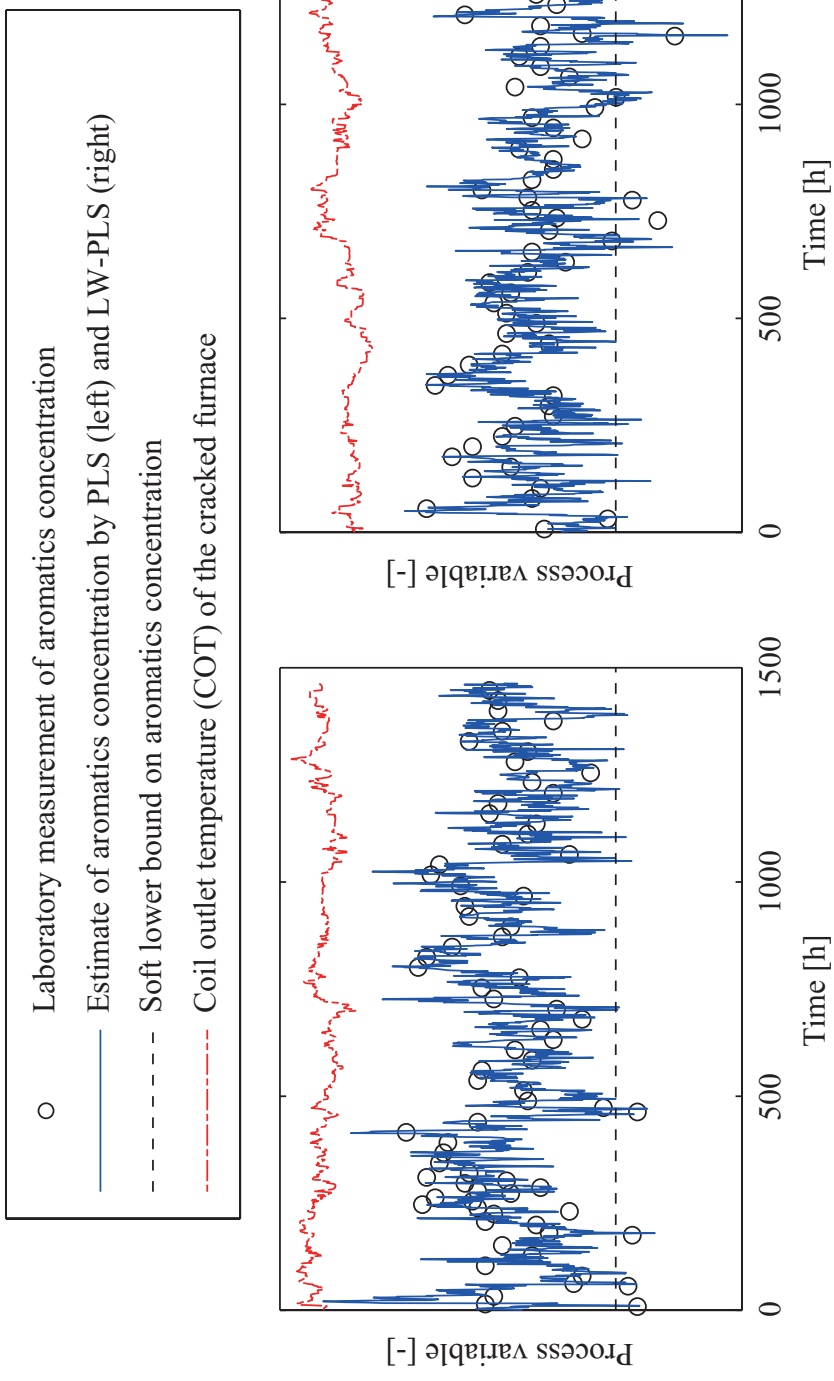

Figure 4.4: Application results of the developed inferential control system integrating MPC and soft-sensors at the ethylene production process in the SDK Oita plant: the aromatics concentration and COT by using PLS (from 1st May to 30th June 2011) (left) and LW-PLS (from 1st October to 30th November 2011) (right)

Table 4.1: Application results of the developed inferential control system integrating MPC and soft-sensors at the ethylene production process in the SDK Oita plant: comparison between PLS-based soft-sensor and LW-PLS-based soft-sensor

	RMSE [%]	$\mathrm{MAD}_{y_{\mathrm{aroma}}}$ [%]	$\sigma_{y_{\mathrm{aroma}}}$ [%]	Mean of COT [%]
PLS	100	100	100	100
LW-PLS	70.4	74.1	93.3	99.6

4.3.3 Purification section for acetyl plant

In the purification section for acetyl plant, the feed is purified using two towers, a degassing tower and a product tower, as shown in Figure 4.5. This process has been operated by using a multivariate MPC controller for more than one and half years; CVs, MVs and DVs are shown in Figure 4.5 and Table 4.2. Twelve measurements and the estimate of the amount of impurity in the product are set as CVs and are controlled by using three MVs: steam flow rate in the degassing tower, reflux flow rate of the product tower and a temperature in the product tower. The amount of impurity in the product has a soft upper constraint and should be kept close to the constraint in order to satisfy the product specification and to reduce operation cost. In other words, short-term violations are acceptable. In this process, neither a soft-sensor nor MPC software was used, and the amount of impurity has been usually measured only once a day thus the control performance had been poor. To solve this problem and minimize the steam flow rates in the degassing tower and the product tower, the developed inferential control system was implemented. Twenty three process variables such as temperature, feed flow rate, differential pressure and steam flow rate were selected as input variables of the soft-sensor on the basis of engineers' process knowledge. The number of latent variables R and the localization parameter φ were determined by cross validation; R and φ were set to five

and 0.8, respectively. The 30 newest samples and core data which consists of 297 samples obtained from 1st May 2010 to 31st October 2011 were used to construct LW-PLS models. The number of newest samples and the core data were selected by trial and error.

Figure 4.6 shows the laboratory measurements of impurity amount y_{im}, the estimates, the constraint and the steam flow rate in the degassing tower before and after LW-PLS was implemented. All variables are scaled in this figure, and the control results before and after the implementation of LW-PLS are compared. In addition, Table 4.3 shows the standard deviation σ of y_{im}, mean absolute deviation (MAD) of y_{im} from the upper bound, mean of the steam flow rate in the degassing tower F_1 and that in the product tower F_2, which is used for control of the temperature in the product tower (MV-T1 in Table 4.2 and Figure 4.5) in a lower level control loop. Here, $\sigma_{y_{\mathrm{im}}}$, $\mathrm{MAD}_{y_{\mathrm{im}}}$ and means of F_1 and F_2 are scaled so that they are 100 before the implementation of LW-PLS. As shown in Figure 4.6 and Table 4.3, the impurity amount has been kept close to the soft upper bound by using the MPC system combined with the LW-PLS-based soft-sensor. In addition, $\sigma_{y_{\mathrm{im}}}$, means of F_1 and F_2 were reduced by 32.5%, 25.4% and 2.8%, respectively, and the operation cost was significantly reduced.

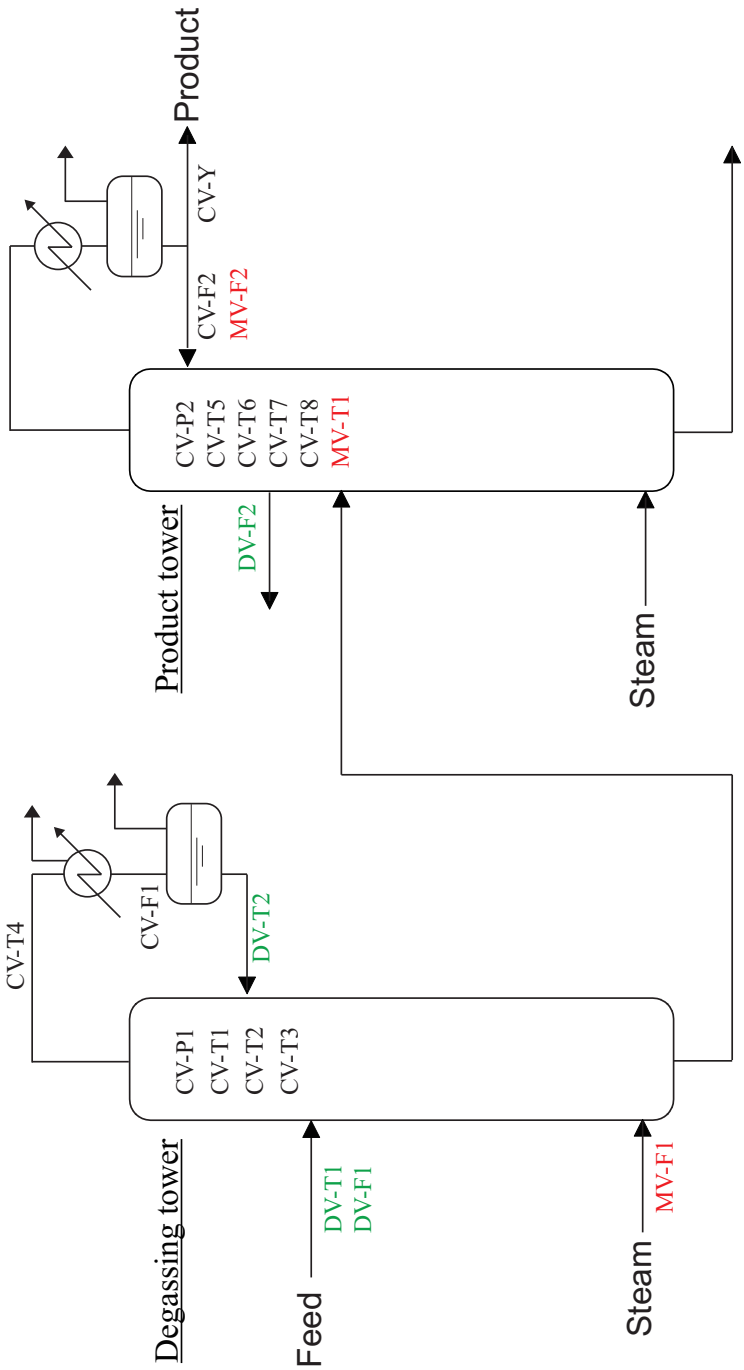

Figure 4.5: Schematic diagram of the purification section for acetyl plant at the SDK Oita plant

Table 4.2: Variables used in MPC in the purification section for acetyl plant (DT and PT denote product tower and degassing tower, respectively)

Tag	Variable name
CV-P1	Differential pressure (DT)
CV-P2	Differential pressure (PT)
CV-T1, 2, 3	Temperatures (DT)
CV-T4	Temperature of the outlet flow from the top (DT)
CV-T5, 6, 7, 8	Temperatures (PT)
CV-F1	Feed flow rate into the reflux drum (DT)
CV-F2	Reflux ratio (PT)
CV-Y	Estimated value of impurity amount (PT)
MV-F1	Steam flow rate (DT)
MV-F2	Reflux flow rate (PT)
MV-T1	Temperature (PT)
DV-T1	Feed temperature (DT)
DV-T2	Temperature of the reflux flow (DT)
DV-F1	Feed flow rate (DT)
DV-F2	Side cut flow rate (PT)

Figure 4.6: Application results of the conventional control system (from 9th April to 9th May 2012) (left) and the developed inferential control system integrating MPC and LW-PLS (from 1st to 31st August 2012) (right) at the purification section for acetyl plant in the SDK Oita plant

Table 4.3: Comparison between the control results by using the conventional control system and the developed inferential control system integrating MPC and LW-PLS at the purification section for acetyl plant in the SDK Oita plant

	$\mathrm{MAD}_{y_{\mathrm{im}}}$ [%]	$\sigma_{y_{\mathrm{im}}}$ [%]	Mean of F_1 [%]	Mean of F_2 [%]
Conventional system	100	100	100	100
MPC & LW-PLS	36.1	67.5	74.6	97.2

4.4 New similarity measure

To construct highly accurate JIT soft-sensors, it is crucial to define the similarity between samples. In the SDK Oita plant, the similarity defined by equations (4.15) and (4.16) has been used because the distance-based similarity is frequently used and the development time is limited; however, the estimation performance could be improved by using another similarity. In this section, to further improve the estimation performance of LW-PLS, a new similarity based on the weighted distance between samples is investigated since past researches [59–61] revealed that it can improve the estimation performance.

$$\omega_n = \exp\left(-\frac{d_n}{\sigma_d \varphi}\right) \tag{4.17}$$

$$d_n^2 = (\boldsymbol{x}_n - \boldsymbol{x}_q)^{\mathrm{T}} \boldsymbol{\Theta} (\boldsymbol{x}_n - \boldsymbol{x}_q) \tag{4.18}$$

$$\boldsymbol{\Theta} = \mathrm{diag}(\theta_1, \theta_2, \cdots, \theta_M) \tag{4.19}$$

Here, $\boldsymbol{\Theta}$ denotes a weighting matrix.

4.4.1 How should weights be determined?

Figure 4.7 shows simple examples, in which a relationship between a local linear model and weights θ_m $(m = 1, 2, 3)$ is illustrated by using very small number of samples. In

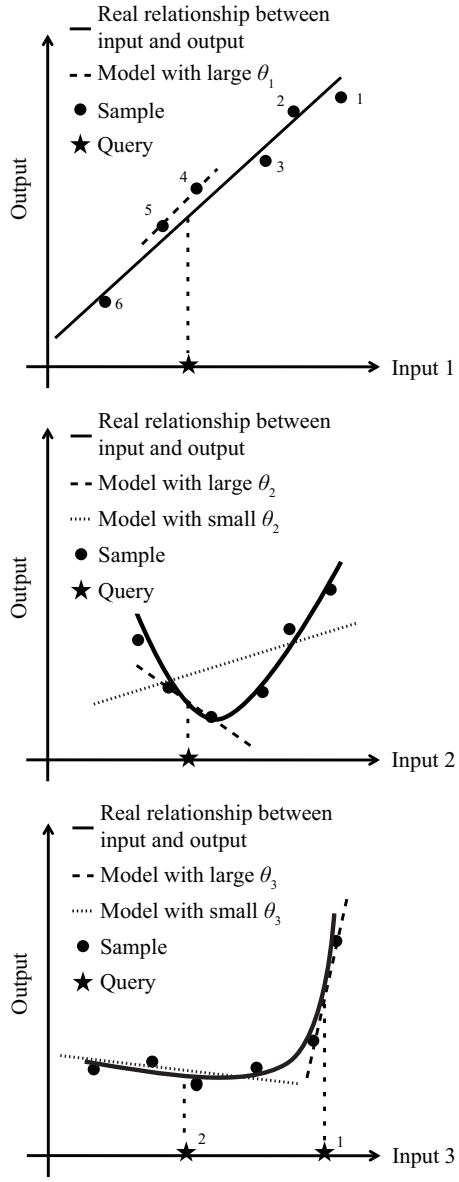

Figure 4.7: The relationship between a local model and weights θ_m ($m = 1, 2, 3$). a case where relationship between an input and an output is linear (top). a case where relationship between an input and an output is nonlinear (middle). a case where the strength of nonlinearity changes depending on the value of an input variable (bottom)

each figure, it is assumed that the values and the weights of the other inputs, and φ are constant. Relationship between input 1 and the output variable is linear as shown in Figure 4.7 (top). Large θ_1 causes overfitting by prioritizing samples 4 and 5; therefore, θ_1 should be small. On the contrary, the relationship between input 2 and the output variable is nonlinear as shown in Figure 4.7 (middle). Thus, θ_2 should be large to cope with nonlinearity between input 2 and the output variable. In addition, the strength of nonlinearity around a query may change depending on the value of the input variable as shown in Figure 4.7 (bottom). In this case, θ_3 should be large for query 1 and small for query 2.

The weights proposed in [59–61], which are based on regression coefficients, do not necessarily correspond to the strength of nonlinearity around a query. For example, a regression coefficient of an input can be large even when the input-output relationship is linear. In such a case, the large weight might cause a deterioration of the estimation performance as shown in Figure 4.7 (top).

4.4.2 Proposed procedure for calculating similarity

Section 4.4.1 revealed that weights of inputs should correspond to strength of nonlinearity between the inputs and an output around a query. In addition, a regression coefficient, i.e. slope in Figure 4.7, significantly changes around a query when the nonlinearity around it is strong. To evaluate the change of the regression coefficient of an input around a query and to determine the weights, the weighted variance of each input's regression coefficients of LW-PLS models is utilized. The similarities between a query and samples in a database are utilized as the weights when the weighted variance is calculated. Since similarity depends on the weight θ, iterative calculation is conducted to derive similarity and θ.

Offline and online calculation procedures of the weights are as follows.

· Offline part

1. Determine the number of latent variables R, the localization parameter φ and the maximum iteration number $I_{\max}$.

2. Set i to 1 and $\theta_{m,i-1}$ to 1 for all m.

3. Regard each of N samples in the database as a query and construct N LW-PLS models by using $\theta_{m,i-1}$.

4. Calculate the variance $V_{m,i}$ of N regression coefficients of the mth input variable $\beta_{nm,i}$, then set $\theta_{m,i}$ to $(V_{m,i})^{\alpha}$. Here, α is a tuning parameter.

$$V_{m,i} = \frac{1}{N-1}\sum_{n=1}^{N}(\beta_{nm} - \bar{\beta}_{*m}) \tag{4.20}$$

$$\bar{\beta}_{*m} = \frac{1}{N}\sum_{n=1}^{N}\beta_{nm} \tag{4.21}$$

5. If $i = I_{\max}$ or the following equation is satisfied for all m, finish the offline calculation. Here, ε_1 is a tolerance.

$$\left|\frac{\theta_{m,i} - \theta_{m,i-1}}{\theta_{m,i-1}}\right| \leq \varepsilon_1 \tag{4.22}$$

6. Set i to $i + 1$ and go to step 3.

In the offline part, θ_m is first set to 1, then θ_m is updated to $(V_{m,i})^{\alpha}$; β_{nm} and V_m are calculated repeatedly until θ_m converges.

· Online part

1. Determine the maximum iteration number $J_{\max}$, and set j to 1 and $\theta_{m,j-1} = \theta_m$ obtained in the offline part.

2. Calculate the similarity $\omega_{n,j-1}$ by using $\theta_{m,j-1}$.

3. Calculate the weighted variance $V_{m,j}$ of β_{nm} obtained in the offline part.

$$V_{m,j} = \sum_{n=1}^{N} \omega_{n,j-1}(\beta_{nm} - \bar{\beta}_m)^2 / \sum_{n=1}^{N} \omega_{n,j-1} \tag{4.23}$$

$$\bar{\beta}_m = \sum_{n=1}^{N} \omega_{n,j-1}\beta_{nm} / \sum_{n=1}^{N} \omega_{n,j-1} \tag{4.24}$$

4. Set $\theta_{m,j}$ to $[(V_{m,j})^{\alpha} + \theta_{m,j-1}]/2$.

5. If $j = J_{\max}$ or the following equation is satisfied for all m, finish the online calculation. Here, ε_2 is a tolerance.

$$\left| \frac{\theta_{m,j} - \theta_{m,j-1}}{\theta_{m,j-1}} \right| \leq \varepsilon_2 \tag{4.25}$$

6. Set j to $j + 1$ and go to step 2.

In the online part, to evaluate the strength of nonlinearity around a query, θ_m is updated by using the weighted variance V_m of β_{nm} obtained in the offline part and the similarity ω_n.

This procedure contains seven parameters to be determined: the number of latent variables R, the localization parameter φ, the tuning parameter α, the maximum iteration number in the offline part $I_{\max}$ and in the online part $J_{\max}$, and the tolerance in the offline part ε_1 and in the online part ε_2. R, φ and α can be determined by applying cross validation to all data or by building and validating models with different datasets, i.e., model construction data and parameter tuning data. The proposed method includes the conventional LW-PLS, which uses normal Euclidean distance since the proposed method becomes the same as the conventional one when $\alpha = 0$. Thus, the estimation accuracy of the proposed LW-PLS model is the same as or better than that of the conventional LW-PLS model when α is tuned properly.

4.5 Numerical example

In this section, the proposed similarity is compared with the conventional ones in two numerical examples. The following four methods are compared.

LW-PLS 1) LW-PLS with $\theta_m = 1$.

LW-PLS 2) LW-PLS with θ_m defined as the absolute value of the mth variable's regression coefficient of a global MLR model [59].

LW-PLS 3) LW-PLS with θ_m defined as the absolute value of the mth variable's regression coefficient of an LW-PLS model constructed by LW-PLS 1 [60].

LW-PLS 4) LW-PLS with θ_m defined by the proposed method.

4.5.1 Problem settings

The following two cases are investigated; in each case, x_m and y are inputs and an output, respectively.

· Case 1

$$w_m \sim N(0, 0.02^2) \ (m = 0, 1, 2, 3) \tag{4.26}$$

$$s_m \sim \mathrm{rand}(-5, 5) \ (m = 1, 2, 3) \tag{4.27}$$

$$x_m = s_m + w_m \ (m = 1, 2, 3) \tag{4.28}$$

$$y = 10s_1 + 5{s_2}^2 + \exp(s_3) + w_0 \tag{4.29}$$

· Case 2

$$w_m \sim N(0, 0.02^2) \ (m = 0, 1, \cdots, 7) \tag{4.30}$$

$$s_m \sim \mathrm{rand}(-5, 5) \ (m = 1, 2, \cdots, 6) \tag{4.31}$$

$$x_m = s_m + w_m \ (m = 1, 2, \cdots, 6) \tag{4.32}$$

$$x_7 = s_6 + w_7 \tag{4.33}$$

$$y = {s_2}^3 + 3s_3 + {s_4}^2 + \exp(s_5) + 3s_6 + w_0 \tag{4.34}$$

Here, rand(a, b) denotes the uniform random distribution in a closed interval [a b], and $N(\mu, \sigma^2)$ denotes the normal distribution whose mean is μ and standard deviation is σ. In both cases, 3000 samples were generated and divided into three groups: samples for model construction (1000 samples), parameter tuning (1000 samples) and model validation (1000 samples). Models were constructed with different values of localization parameter φ, the number of latent variables R, and α, by using samples for model construction. Then, the estimation errors were calculated by using samples for parameter tuning, and the set of parameters that minimized the estimation error was selected. The search range of φ, R and α is [0.01, 0.03, $\cdots$, 0.09], [1, 2, 3] and [0.01, 0.03, $\cdots$, 0.09], respectively. The appropriate search range of the parameters depends on the situation; therefore, it is recommended to make the search range wide enough in order to get the optimal parameters. In LW-PLS 4, tolerances ε_1 and ε_2 are set to 0.01. Both of the maximum iteration numbers $I_{\max}$ and $J_{\max}$ are 30.

4.5.2 Results and discussion

Table 4.4 shows the selected parameters and root mean square error for validation samples (RMSEV). The proposed method achieved the minimum RMSEV in both cases and was considerably superior to the conventional methods. Figure 4.8 shows the relation-

ship between RMSE for parameter tuning samples (RMSET) and φ when the proposed method is applied to case 1 (R = 3). RMSET was large when φ was too small or too large. Overfitting occurred when φ was too small, and models were unable to cope with nonlinearity between input and output variables when φ was too large. Tables 4.5 and 4.6 show θ_m when $\boldsymbol{x}_q = [0, 0, 3]^{\mathrm{T}}$ and $\boldsymbol{x}_q = [0, 0, -3]^{\mathrm{T}}$ in case 1, respectively. Figure 4.9 shows the change of θ_m in the online part of the weights calculation procedure. Here, θ_m is normalized so that the sum of θ_m be 1 in LW-PLS 2, 3 and 4. In case 1, the relationship between x_1 and the output is linear, therefore, θ_1 should be 0. When $\boldsymbol{x}_q = [0, 0, 3]^{\mathrm{T}}$, θ_3 should be larger than θ_2 because x_3 has stronger nonlinearity around the query than x_2, i.e.

$$\mathrm{abs}\left(\left.\frac{\partial^2 y}{\partial {x_2}^2}\right|_{\boldsymbol{x}=[0,0,3]^{\mathrm{T}}}\right) < \mathrm{abs}\left(\left.\frac{\partial^2 y}{\partial {x_3}^2}\right|_{\boldsymbol{x}=[0,0,3]^{\mathrm{T}}}\right) \tag{4.35}$$

where abs(a) denotes the absolute value of a. On the other hand, when $\boldsymbol{x}_q = [0, 0, -3]^{\mathrm{T}}$, θ_2 should be larger than θ_3 because

$$\mathrm{abs}\left(\left.\frac{\partial^2 y}{\partial {x_2}^2}\right|_{\boldsymbol{x}=[0,0,-3]^{\mathrm{T}}}\right) > \mathrm{abs}\left(\left.\frac{\partial^2 y}{\partial {x_3}^2}\right|_{\boldsymbol{x}=[0,0,-3]^{\mathrm{T}}}\right). \tag{4.36}$$

The proposed method derived appropriate θ for both queries. On the other hand, LW-PLS 2 and 3 were not able to derive θ properly: θ_1 is large since the regression coefficient of x_1 is large though x_1 dose not have nonlinear effect on the output variable. This is the reason why the proposed method achieved the best performance in the four methods.

Table 4.4: Selected parameters and RMSEV in numerical examples

Case	Method	R	φ	α	RMSEV
1	LW-PLS 1	3	0.05	-	3.84
	LW-PLS 2	3	0.05	-	5.29
	LW-PLS 3	3	0.05	-	4.69
	LW-PLS 4	3	0.03	0.8	1.59
2	LW-PLS 1	6	0.21	-	18.93
	LW-PLS 2	6	0.09	-	16.53
	LW-PLS 3	5	0.21	-	21.17
	LW-PLS 4	6	0.06	0.8	5.31

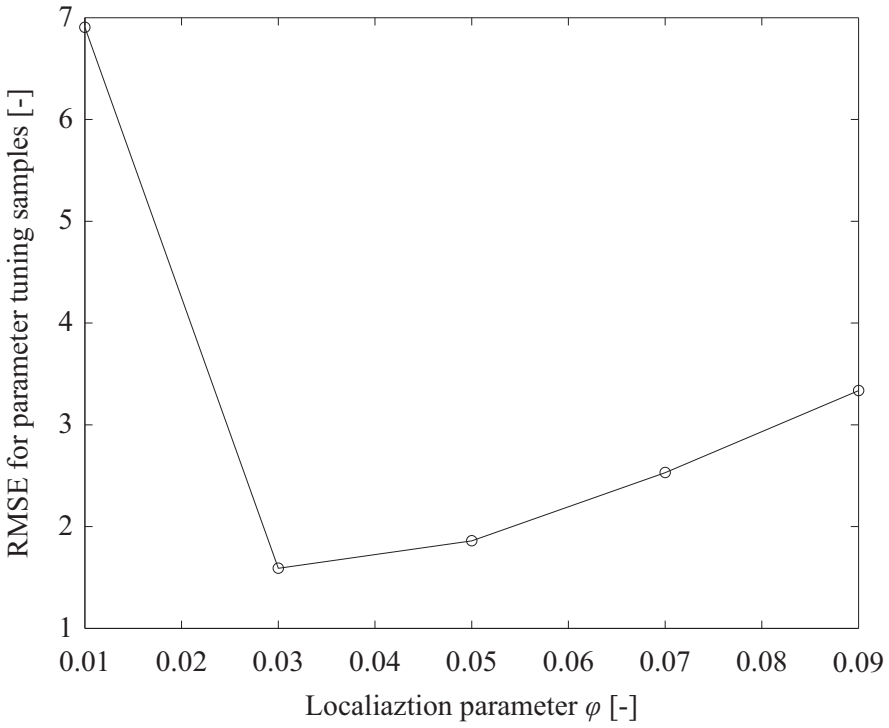

Figure 4.8: The relationship between RMSE for parameter tuning samples and the localization parameter φ when the proposed method is applied in case 1 (R = 3)

Table 4.5: Derived weights of $\boldsymbol{x}_q = [0, 0, 3]^{\mathrm{T}}$ in case 1

Method	θ_1	θ_2	θ_3
LW-PLS 1	1.00	1.00	1.00
LW-PLS 2	0.53	0.08	0.39
LW-PLS 3	0.32	0.02	0.66
LW-PLS 4	0.00	0.32	0.68

Table 4.6: Derived weights of $\boldsymbol{x}_q = [0, 0, -3]^{\mathrm{T}}$ in case 1

Method	θ_1	θ_2	θ_3
LW-PLS 1	1.00	1.00	1.00
LW-PLS 2	0.53	0.08	0.39
LW-PLS 3	0.57	0.34	0.09
LW-PLS 4	0.01	0.81	0.18

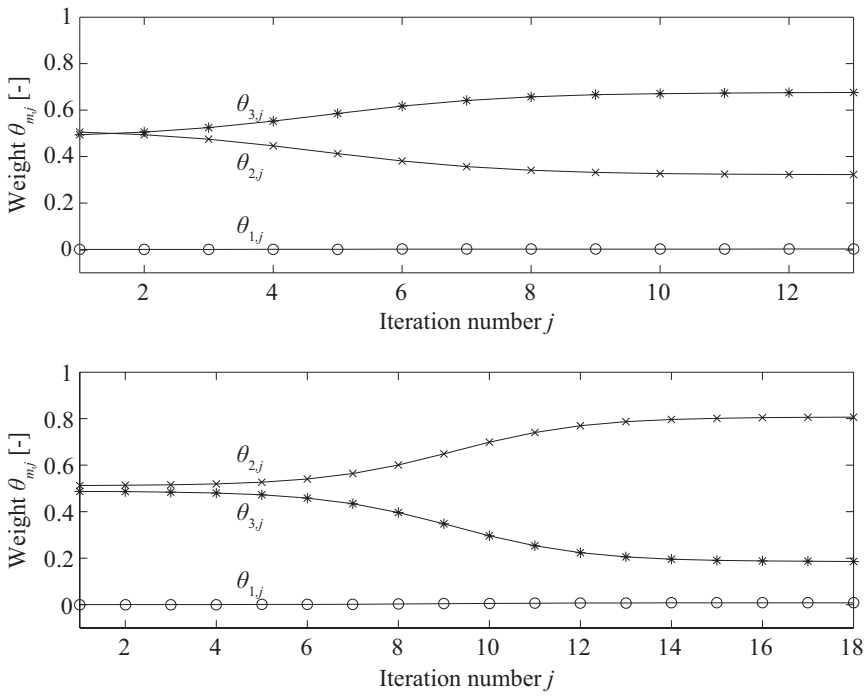

Figure 4.9: Change of θ_m in the online part of the weights calculation procedure in case 1. $\boldsymbol{x}_q = [0, 0, 3]^{\mathrm{T}}$ (top). $\boldsymbol{x}_q = [0, 0, -3]^{\mathrm{T}}$ (bottom)

4.6 Application to distillation process

In this section, an application result of the proposed similarity to an industrial distillation process is reported. A soft-sensor for estimating the aromatics concentration was constructed in order to realize highly efficient operation of the CGL fractionator of the ethylene production process at the SDK Oita plant in Japan. In this case study, linear PLS, LW-PLS 1, 2, 3 and 4 explained in the previous section were compared. The operation data obtained from January 1, 2010 to August 4, 2011 were stored in the database. The tuning parameters were determined using these data. The search range of φ, R and α is [0.2, 0.4, $\cdots$, 2, 2.5, 3.0, $\cdots$,10], [1, 2, $\cdots$,9], [0.2, 0.4, $\cdots$, 2.0], respectively. Then, the aromatics concentration was estimated for the operation data obtained from August 6, 2011 to December 31, 2011.

4.6.1 Results and discussion

Table 4.7 shows the selected parameters and RMSEV. In LW-PLS 4, tolerances ε_1 and ε_2 are 0.01. The maximum iteration numbers $I_{\max}$ and $J_{\max}$ are 20 and 30, respectively. The average calculation time of output estimation for each query was 4.8 msec when Intel® Core™ i7-2620M (2.7 GHz×2) and 8 GB RAM were used.

In this process, the output variable (aromatics concentration) is measured to one place of decimal, thus, the differences of RMSEs between linear PLS and LW-PLS 1, and between LW-PLS 2, 3 and 4 are not significant. The reason why LW-PLS 2, 3 and 4 derived the better result than the other methods might be that the strength of nonlinear effect of each input on the output is different. Table 4.8 shows the maximum, mean and minimum values, and standard deviation of the mth weight θ_m when LW-PLS 4 is applied. Here, $\theta_1, \theta_2, \cdots, \theta_9$ for each query are normalized so that their sum becomes 1. θ_1 is the largest and the nonlinear effect of input 1 on the output is expected to be strong. In addition, the strength of nonlinear effect of each input on the output does not seem to

Table 4.7: Selected parameters and RMSEV in a case study of the CGL fractionator

Method	R	φ	α	RMSEV
Linear PLS	2	-	-	1.20
LW-PLS 1	2	6.5	-	1.15
LW-PLS 2	2	1.0	-	0.99
LW-PLS 3	2	1.0	-	0.98
LW-PLS 4	2	1.4	1.2	1.03

Table 4.8: Changes of weights in a case study of the CGL fractionator when LW-PLS 4 is applied

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9
Maximum value	0.32	0.12	0.13	0.09	0.12	0.10	0.09	0.10	0.12
Mean value	0.26	0.09	0.11	0.08	0.09	0.09	0.09	0.08	0.11
Minimum value	0.21	0.07	0.10	0.07	0.08	0.08	0.08	0.08	0.09
Standard deviation	0.03	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.01

depend on the value of input variables since the standard deviations of the weights are small. This could be the reason why RMSEs of LW-PLS 2, 3 and 4 are similar.

4.7 Conclusions

This chapter showed successful applications of inferential control systems based on locally weighted partial least squares (LW-PLS) and commercial model predictive control (MPC) software. The developed systems reduced the cost and time of implementation of soft-sensors, and have been stably working for more than two years at the Showa Denko

K.K. (SDK) Oita plant in Japan. The operation cost and environmental burden have been significantly reduced. In the cracked gasoline (CGL) fractionator, for example, about 0.6% operation cost was cut successfully by decreasing the coil outlet temperature (COT) of the cracking furnace. In the purification section for acetyl plant, the operation cost was significantly reduced by keeping the impurity amount in the product close to the upper bound. Furthermore, the implementation of the developed inferential control system has contributed toward reducing the burden of model maintenance, which was recognized as the most serious problem in practice. The inferential control system combining MPC and the LW-PLS-based soft-sensor is now spreading as a standard process control method to other processes of SDK.

In addition, to further improve the accuracy of LW-PLS models, an adaptive similarity measure was proposed. In the proposed method, weights of input variables are determined through iterative calculation by using the weighted variance of the regression coefficients. The results of the case studies showed that the proposed method could adaptively derive the appropriate weights and more accurate models than the conventional methods in numerical examples. Furthermore, root mean square error was improved by 11.3% by using the proposed method compared to LW-PLS in which conventional similarity based on the Euclidean distance without weights is used. These results clearly demonstrate the usefulness of the proposed method, which uses newly defined similarity based on the weighted Euclidean distance.

Chapter 5

Conclusions

In this thesis, the statistical modeling methods for real-time estimation of product quality were investigated to improve the efficiency of industrial processes. In soft-sensor design procedure, this research focused on input variable scaling, input variable selection, model construction and maintenance, and developed the novel methods that improve the estimation accuracy of soft-sensors and reduce the cost for soft-sensor design.

Chapter 2 investigated the input variable scaling methods. It was revealed that input variable scaling can have significant effect on estimation accuracy, and conventional input scaling methods, which assume all input variables are equally important, is not always optimal. To solve this problem, two novel input scaling methods that can evaluate the importance of input variables, were proposed. One method statistically derives the input scaling factors. The other one utilizes spectroscopic data of a material whose content is an estimation target. The proposed methods successfully improved the estimation accuracy in case studies in pharmaceutical and distillation processes.

In addition to appropriately define the input scaling factors, selection of input variable is important for soft-sensor design. Though many input variable selection methods have been proposed, trial and error is unavoidable since they do not take into account properties of the target processes, and optimal methods depend on each situation. Chap-

ter 3 proposed an input variable selection method for the batch processes. The proposed method takes into account the properties of batch processes, thus the estimation accuracy can be improved. In a case study in a pharmaceutical process, the estimation error was reduced by 28.7 and 45.5% in root mean square error of validation (RMSEV) compared to variable influence on projection (VIP) and a method based on engineers' process knowledge, respectively.

Chapter 4 dealt with the problem of performance deterioration of soft-sensor after implementation, which is recognized as the most serious problem in practice. To solve this problem, locally weighted partial least squares (LW-PLS), i.e. a JIT modeling method, was implemented with model predictive control (MPC) to chemical processes. The developed systems have been stably working for more than two years without performance deterioration, and the operation cost and model maintenance burden have been significantly reduced. In addition, a new similarity measure was proposed to enhance the estimation accuracy of LW-PLS models. The proposed similarity can be adaptively defined on the basis of the condition of target process. By using the proposed method, the estimation accuracy was improved by 11.3% compared to LW-PLS in which conventional similarity was used.

Although the methods developed in this research can contribute to the improved estimation performance of soft-sensors, they were individually applied and an integrated standard soft-sensor design method has to be developed to make soft-sensors more widespread and to maximize the benefit of soft-sensor. To realize this, the following problems should be tackled.

1. How to select samples for soft-sensor design?

2. How to clarify the reason when the estimation performance of a soft-senor is not satisfactory, and how to improve it?

First problem is quite important since the performance of the soft-sensors significantly

depends on the samples used for soft-sensor design. For example, the existence of abnormal samples can deteriorate the performance of input variable selection and model construction methods, since most of them assume that the abnormal samples are not contained in the database. In addition, detection of abnormal samples is not enough since the amount and the position of the sample in the input variable space affect on the estimation performance. Thus a method for evaluating the estimation performance based on the amount and the position of the sample is required in addition to abnormal sample detection.

Second problem is also important to make soft-sensors more familiar to the people who are not the expert of soft-sensor design. However, it is very difficult even for experienced people to clarify the reason of bad estimation performance and improve it . To solve this problem, all methods used in each step of soft-sensor design procedure must be developed concurrently. In addition, the meaning and the results of each method should be easy to interpret as the input variable selection method proposed in chapter 2.

Data analysis technique is becoming much more popular than ever accompanied with increasing power of computers not only in engineering but also in many fields such as business science and medicine. Thus, the demand for developing new data analysis techniques would increase and active researches are encouraged. In addition, it is quite important to collaborate with people in other fields to get more benefit from data analysis, since knowledge about statistics and the target systems are essential.

Bibliography

[1] M. Kano and Y. Nakagawa. Data-based process monitoring, process control, and quality improvement: recent developments and applications in steel industry. *Comput. and Chem. Eng.*, 32(1-2):12–24, 2008.

[2] P. Kadlec, B. Gabrys, and S. Strandt. Data-driven soft sensors in the process industry. *Comput. and Chem. Eng.*, 33:795–814, 2009.

[3] M. Kano and K. Fujiwara. Virtual sensing technology in process industries: trends and challenges revealed by recent industrial applications. *J. Chem. Eng. Jpn.*, 46:1–17, 2013.

[4] R. A. van den Berg, H. C. J Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf. Centering, scaling, and transformations improving the biological information content of metabolomics data. *BMC Genomics*, 7:142, 2006.

[5] R. Todeschini, V. Consonni, and A. Maiocchi. The k correlation index theory development and its application in chemometrics. *Chemom. Intell. Lab. Syst.*, 46:13–29, 1999.

[6] W. A. Shewhart. Statistical method from the viewpoint of quality control. *W. Edwards Swillia Deming*, 1939.

[7] F. R. Hampel. The influence curve and its role in robust estimation. *J. Am. Stat. Assoc.*, 69:383–393, 1974.

[8] J. V. Kresta, J. F. MacGregor, and T. E. Marlin. Multivariate statistical monitoring of process operating performance. *Can. J. Chem. Eng.*, 69:35–47, 1991.

[9] H. Kamohara, A. Takinami, M. Takeda, M. Kano, S. Hasebe, and I. Hashimoto. Product quality estimation and operating condition monitoring for industrial ethylene fractionator. *J. Chem. Eng. Jpn.*, 37(3):422–428, 2004.

[10] M. Kano, S. Tanaka, S. Hasebe, I. Hashimoto, and H. Ohno. Monitoring independent components for fault detection. *AIChE J.*, 49(4):969–976, 2003.

[11] S. J. Qin. Statistical process monitoring: basics and beyond. *J. Chemom.*, 17:480–502, 2003.

[12] E. I. George. The variable selection problem. *J. Am. Stat. Assoc.*, 95(452):1304–1308, 2000.

[13] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, 19(6):716–723, 1974.

[14] C. L. Mallows. Some comments on c_P. *Technometrics*, 15(4):661–675, 1973.

[15] M. Arakawa, Y. Yamashita, and K. Funatsu. Genetic algorithm-based wavelength selection method for spectral calibration. *J. Chemom.*, 25(1):10–19, 2011.

[16] D. Jouen-Rimbauda and D. L. Massart. Genetic algorithms as a tool for wavelength selection in multivariate calibration. *Anal. Chem.*, 67(23):4295–4301, 1995.

[17] S. Wold, M. Sjöström, and L. Eriksson. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.*, 58(2):109 – 130, 2001.

[18] V. Centner, D.-L. Massart, O. E. Noord, S. Jong, B. M. Vandeginste, and C. Sterna. Elimination of uninformative variables for multivariate calibration. *Anal. Chem.*, 68:3851 – 3858, 1996.

[19] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[20] I.-G. Chong and C.-H. Jun. Performance of some variable selection methods when multicollinearity is present. *Chemom. Intell. Lab. Syst.*, 78:103–112, 2005.

[21] Z. Xiaobo, Z. Jiewen, M. J. W. Povey, M. Holmes, and M. Hanpin. Variables selection methods in near-infrared spectroscopy. *Anal. Chim. Acta*, 667:14–32, 2010.

[22] K. Fujiwara, H. Sawada, and M. Kano. Input variable selection for pls modeling using nearest correlation spectral clustering. *Chemom. Intell. Lab. Syst.*, 118:109–119, 2012.

[23] K. Fujiwara, M. Kano, S. Hasebe, and A. Takinami. Soft-sensor development using correlation-based just-in-time modeling. *AIChE J.*, 55:1754–1765, 2009.

[24] K. Fujiwara, M. Kano, and S. Hasebe. Development of correlation-based clustering method and its application to software sensing. *Chemom. Intell. Lab. Syst.*, 101:130–138, 2010.

[25] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø. A review of variable selection methods in partial least squares regression. *Chemom. Intell. Lab. Syst.*, 118:62–69, 2012.

[26] Z. Hui. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.*, 101(476):1418–1429, 2006.

[27] R. M. Balabin and S. V. Smirnov. Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data. *Anal. Chim. Acta*, 692:63–72, 2011.

[28] J. J. Moes, M. M. Ruijken, E. Gout, H. W. Frijlink, and M. I. Ugwoke. Application of process analytical technology in tablet process development using nir spectroscopy: Blend uniformity, content uniformity and coating thickness measurements. *Int. J. Pharm.*, 357(1-2):108–118, 2008.

[29] H. Berthiaux, V. Mosorov, L. Tomczak, C. Gatumel, and J. F. Demeyre. Principal component analysis for characterising homogeneity in powder mixing using image processing techniques. *Chem. Eng. Process.*, 45(5):397–403, 2006.

[30] R. P. Cogdill, C. A. Anderson, M. Delgado-Lopez, D. Molseed, R. Chisholm, R. Bolton, T. Herkert, A. M. Afnan, and J. K. Drennen III. Process analytical technology case study part I: Feasibility studies for quantitative near-infrared method development. *AAPS PharmSciTech*, 6(2):262–272, 2005.

[31] H. Wu, M. Tawakkul, M. White, and M. A. Khan. Quality-by-design (QbD): An integrated multivariate approach for the component quantification in powder blends. *Int. J. Pharm.*, 372(1-2):39–48, 2009.

[32] W. Li and G. D. Worosila. Quantitation of active pharmaceutical ingredients and excipients in powder blends using designed multivariate calibration models by near-infrared spectroscopy. *Int. J. Pharm.*, 295(1-2):213–219, 2005.

[33] O. Berntsson, L. G. Danielsson, B. Lagerholm, and S. Folestad. Quantitative in-line monitoring of powder blending by near infrared reflection spectroscopy. *Powder Technol.*, 123(2-3):185–193, 2002.

[34] Y. Sulub, B. Wabuyele, P. Gargiulo, J. Pazdan, J. Cheney, J. Berry, A. Gupta, R. Shah, H. Wu, and M. Khan. Real-time on-line blend uniformity monitoring using near-infrared reflectance spectrometry: A noninvasive off-line calibration approach. *J. Pharm. Biomed. Anal.*, 49(1):48–54, 2009.

[35] S. Virtanen, O. Antikainen, and J. Yliruusi. Uniformity of poorly miscible powders determined by near infrared spectroscopy. *Int. J. Pharm.*, 345(1-2):108–115, 2007.

[36] G. Andersson, P. Kaufmann, and L. Renberg. Non-linear modelling with a coupled neural network - PLS regression system. *J. Chemom.*, 10:605–614, 1996.

[37] G. Baffi, E. B. Martin, and A. J. Morris. Non-linear projection to latent structures revisited (the neural network PLS algorithm). *Comput. Chem. Eng.*, 23:1293–1307, 1999.

[38] S. J. Qin and T. J. McAvovy. Nonlinear PLS modeling using neural networks. *Comput. Chem. Eng.*, 16:379–391, 1992.

[39] D. Pérez-Marin, A. Garrido-Varo, J. E. Guerrero, and J. C. Gutiérrez-Estrada. Use of artifcial neural networks in near-infrared reflectance spectroscopy calibrations for predicting the inclusion percentages of wheat and sunflower meal in compound feedingstuffs. *Appl. Spectrosc.*, 60:1062–1069, 2006.

[40] V. R. Nadadoor, H. Siegler, S. L. Shah, W. C. McCaffrey, and A. Ben-Zvi. Online sensor for monitoring a microalgal bioreactor system using support vector regression vector regression. *Chemom. Intell. Lab. Syst.*, 44:2101–2105, 2012.

[41] I. Barman, C. R. Kong, N. C. Dingari, R. R. Dasari, and M. S. Feld. Development of robust calibration models using support vector machines for spectroscopic monitoring of blood glucose. *Chemom. Intell. Lab. Syst.*, 82:9719–9726, 2012.

[42] D. E. Lee, J. H. Song, S. O. Song, and E. S. Yoon. Weighted support vector machine for quality estimation in the polymerization process. *Ind. Eng. Chem. Res.*, 44:2101–2105, 2005.

[43] S. Wold, N. K. Wold, and B. Skagerberg. Nonlinear PLS modeling. *Chemom. Intell. Lab. Syst.*, 7:53–65, 1989.

[44] G. Robertsson. Contributions to the problem of approximation of non-linear data with linear pls in an absorption spectroscopic context. *Chemom. Intell. Lab. Syst.*, 47:99–106, 1999.

[45] A. I. Abdel-Rahmana and G. J. Lim. A nonlinear partial least squares algorithm using quadratic fuzzy inference system. *J. Chemom.*, 23:530–537, 2009.

[46] P. Kadlec, B. Gabrys, and S. Strandt. Data-driven soft sensors in the process industry. *Comput. and Chem. Eng.*, 33:795–814, 2009.

[47] S. Y. Chang, E. H. Baughman, and B. C. McIntosh. Implementation of locally weighted regression to maintain calibrations on FT-NIR analyzers for industrial processes. *Appl. Spectrosc.*, 55:1199–1206, 2001.

[48] P. Kadlec, R. Grbić, and B. Gabrys. Review of adaptation mechanisms for data-driven soft sensors. *Comput. and Chem. Eng.*, 35:1–24, 2011.

[49] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.*, 74:829–836, 1979.

[50] A. T. Walden and P. Prescott. Identification of trends in annual maximum sea levels using robust locally weighted regression. *Estuar. Coast. Shelf. S.*, 16:17–26, 1983.

[51] W. S. Cleveland and S. J. Devlin. Locally weighted regression: An approach to regression analysis by local fitting. *J. Am. Stat. Assoc.*, 83:596–610, 1988.

[52] T. Naes and T. Isaksson. Locally weighted regression and scatter correction for near-infrared reflectance data. *Anal. Chem.*, 62:664–673, 1990.

[53] V. Centner and D. L. Massart. Optimization in locally weighted regression. *Anal. Chem.*, 70:4206–4211, 1998.

[54] H. H. Leung, Y. S. Huang, and C. X. Cao. Locally weighted regression for desulphurisation intelligent decision system modeling. *Simulat. Model. Pract. Theor.*, 12:413–423, 2004.

[55] Z. Ge and Z. Song. A comparative study of just-in-time-learning based methods for online soft sensor modeling. *Chemom. Intell. Lab. Syst.*, 104:306–317, 2010.

[56] C. Cheng and M. S. Chiu. A new data-based methodology for nonlinear process modeling. *Chem. Eng. Sci.*, 59:2801–2810, 2004.

[57] Z. Ge and Z. Song. Online monitoring of nonlinear multiple mode processes based on adaptive local model approach. *Control Eng. Pract.*, 16:1427–1437, 2008.

[58] Z. Y. Wang, T. Isaksson, and B. R. Kowalski. New approach for distance measurement in locally weighted regression. *Anal. Chem.*, 66:249–260, 1994.

[59] H. Shigemori, M. Kano, and S. Hasebe. Optimum quality design system for steel products through locally weighted regression model. *J. Proc. Cont.*, 21(2):293–301, 2011.

[60] S. Kim, M. Kano, H. Nakagawa, and S. Hasebe. Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection. *Int. J. Pharm.*, 421:269–274, 2011.

[61] H. Nakagawa, T. Tajima, M. Kano, S. Kim, S. Hasebe, T. Suzuki, and H. Nakagami. Evaluation of infrared-reflection absorption spectroscopy measurement and

locally weighted partial least-squares for rapid analysis of residual drug substances in cleaning processes. *Anal. Chem.*, 84:3820–3826, 2012.

[62] W.-M. Wu, F.-T. Cheng, and F.-W. Kong. Dynamic-moving-window scheme for virtual-metrology model refreshing. *IEEE Trans. Semicond. Manuf.*, 25(2):238–246, 2012.

[63] R. Bro and A. K. Smilde. Centering and scaling in component analysis. *J. Chemom.*, 17:16–33, 2003.

[64] H. C. Keun, T. M. D. Ebbels, H. Antti, M. E. Bollard, O. Beckonert, E. Holmes, J. C. Lindon, and J. K. Nicholson. Improved analysis of multivariate data by variable stability scaling application to nmr-based metabolic profiling. *Anal. Chim. Acta*, 490:265–276, 2003.

[65] I. Kuzmanovski, M. Novi, and M. Trpkovskaa. Automatic adjustment of the relative importance of different input variables for optimization of counter-propagation artificial neural networks. *Anal. Chim. Acta*, 642:142–147, 2009.

[66] H. Martens, M. Hoy, B. M. Wise, R. Bro, and P. B. Brockhoff. Pre-whitening of data by covariance-weighted pre-processing. *J. Chemom.*, 17:153–165, 2003.

[67] FDA. Pharmaceutical cGMPs for the 21st century - A risk-based approach final report. 2004.

[68] ICH. ICH harmonised tripartite guideline - Pharmaceutical development Q8 (R2). 2005.

[69] ICH. ICH harmonised tripartite guideline - Quality risk management Q9. 2005.

[70] ICH. ICH harmonised tripartite guideline - Pharmaceutical quality system Q10. 2008.

[71] Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, and N. Jent. A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *J. Pharm. Biomed. Anal.*, 44(3):683–700, 2007.

[72] G. Reich. Near-infrared spectroscopy and imaging: Basic principles and pharmaceutical applications. *Adv. Drug Delivery Rev.*, 57(8):1109–1143, 2005.

[73] A. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.*, 36(8):1627–1639, 1964.

[74] R. J. Barnes, M. S. Dhanoa, and S. J. Lister. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc.*, 43(5):772–777, 1989.

[75] S. Schaal, C. G. Atkeson, and S. Vijayakumar. Scalable techniques from nonparametric statistics for real time robot learning. *Appl. Intell.*, 17(60):49–60, 2002.

Acknowledgements

I am deeply grateful to Prof. Shinji Hasebe and Prof. Manabu Kano for their helpful and encouraging comments. Without their dedicated support, this work could not be completed. In the future, I wish to be such a nice teacher and to return the courtesy.

Special thanks to Prof. Masahiro Oshima and Prof. Minoru Miyahara for giving insightful comments and suggestion as reviewers of this thesis.

I would like to the appreciation to Prof. Biao Huang in University of Alberta for giving me an opportunity to conduct the research abroad. It was impressive experience for me.

This study owes a huge debt to co-researchers in industries, Mr. Hiroshi Nakagawa, Mr. Takeshi Seki and Mr. Akitoshi Takinami, for their suggestive comments and providing the real process data. It is great pleasure for me that the developed methods have been adopted in the real processes.

I am grateful to Assist. Prof. Osamu Tonomura for his operational management of our laboratory. I enjoyed wonderful environment for research.

I would also like to thank Ms. Yoshiko Nakanishi, a secretary of my laboratory, for her kind attitude and support for office work.

In addition, it is great honor to have seniors, peers and juniors who share good memories.

Finally, my deepest gratitude gose to my family.

List of Publications

Journal paper

1. S. Kim, M. Kano, H. Nakagawa, and S. Hasebe. Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection. *Int. J. Pharm.*, 421(2): 269-274, (2011)

2. S. Kim, R. Okajima, M. Kano, and S. Hasebe. Development of soft-sensor using locally weighted PLS with adaptive similarity measure. *Chemom. Intell. Lab. Syst.*, 124: 43-49, (2013)

3. S. Kim, M. Kano, S. Hasebe, A. Takinami, and T. Seki. Long-term industrial applications of inferential control based on just-in-time soft-sensors: economical impact and challenges. *Ind. Eng. Chem. Res.*, 52(35): 12346-12356, (2013)

4. S. Kim, M. Kano, H. Nakagawa, and S. Hasebe. Input variable scaling for statistical modeling. *J. Chemom.*, (submitted)

Conference proceeding

1. S. Kim, M. Kano, H. Nakagawa, and S. Hasebe. Estimation of active pharmaceutical ingredients content in blending process for drug products manufacturing. *AIChE Annual Meeting*, 445d, Salt Lake City, US, Nov. (2010)

2. S. Kim, M. Kano, H. Nakagawa, and S. Hasebe. Active pharmaceutical ingredients content estimation using locally weighted partial least squares and statistical wavelength selection. *IFPAC Annual Meeting*, I-028, Baltimore, US, Jan. (2011)

3. R. Okajima, S. Kim, M. Kano, and S. Hasebe. Selection of similarity measure for locally weighted partial least squares regression. *AIChE Annual Meeting*, 669e, Minneapolis, US, Oct. (2011)

4. S. Kim, M. Kano, H. Nakagawa, and S. Hasebe. Estimation of active pharmaceutical ingredient content using locally weighted partial least squares. *Foundations of Computer-Aided Process Operations 2012/Chemical Process Control VIII*, Contributed paper #6, Savannah, US, Jan. (2012)

5. S. Kim, R. Okajima, M. Kano, H. Nakagawa, and S. Hasebe. Soft-sensor using locally weighted PLS with adaptive similarity measure and its application to pharmaceutical process. *9th World Congress of Chemical Engineering incorporating 15th Asian Pacific Confederation of Chemical Engineering Congress*, MoP-T1-119, Seoul, Korea, Aug. (2013)